Empowering A New Plant Biology

Discovery Environment Manual

©2011 iPlant Collaborative - The iPlant Collaborative is funded by a grant from the National Science Foundation Plant Cyberinfrastructure Program (#EF-0735191).

Discovery Environment Manual

1 Getting Started

1.1 Accessing the Discovery Environment 5

1.2 Discovery Environment Overview 7

1.3 Manage Data 10

1.4 Perform Analyses 17

1.5 Viewing and Deleting Notifications 23

2 Analyses

2.1 Ancestral Character Estimation (ACE) Overview 26

2.2 Continuous Ancestral Character Estimation (CACE) 27

2.3 Discrete Ancestral Character Estimation (DACE) 32

2.4 Burrows-Wheeler Aligner Single End Reads 38

2.5 Burrows-Wheeler Aligner Paired End Reads 41

2.6 Cufflinks Transcript Quantification 44

2.7 FASTX Analyses Overview 49

2.8 FASTX Barcode Splitter (Single End) 50

2.9 FASTX Clipper 54

2.10 FASTX Quality Filter 59

2.11 FASTQ Quality Rescaler 62

2.12 FASTX Trimmer 65

2.13 Find SNPs Overview 68

2.14 Find SNPs 69

2.15 Independent Contrasts Overview 76

2.16 Independent Contrasts 77

2.17 Taxonomic Name Resolution Service (TNRS) Demo 83

2.18 TopHat Single End for Illumina 89

2.19 TopHat Paired End for Illumina 95

3 Tools

3.1 Tools Overview 103

3.2 Analysis of Phylogenetics and Evolution (ape) 104

3.3 Burrows-Wheeler Aligner (BWA) 105

3.4 Contrast 106

3.5 Cufflinks 107

3.6 FASTX-Toolkit 108

3.7 R Language and Environment 109

3.8 SAMtools 110

3.9 TopHat 111

3.10 Tree Reconciliation Demo 112

4 Reference

4.1 Discovery Environment 0.3.0 Release Notes 117

4.2 Tool Integration 129

4.3 Creating a New Analysis in the Discovery Environment 130

4.4 TestData folder contents 131

Getting Started

Accessing the Discovery Environment

Account request and creation

Create an account from the iPlant Collaborative website at http://www.iplantcollaborative.org by moving your mouse cursor over the Discovery Environment tab and selecting Request Access to DE from the drop-down menu.

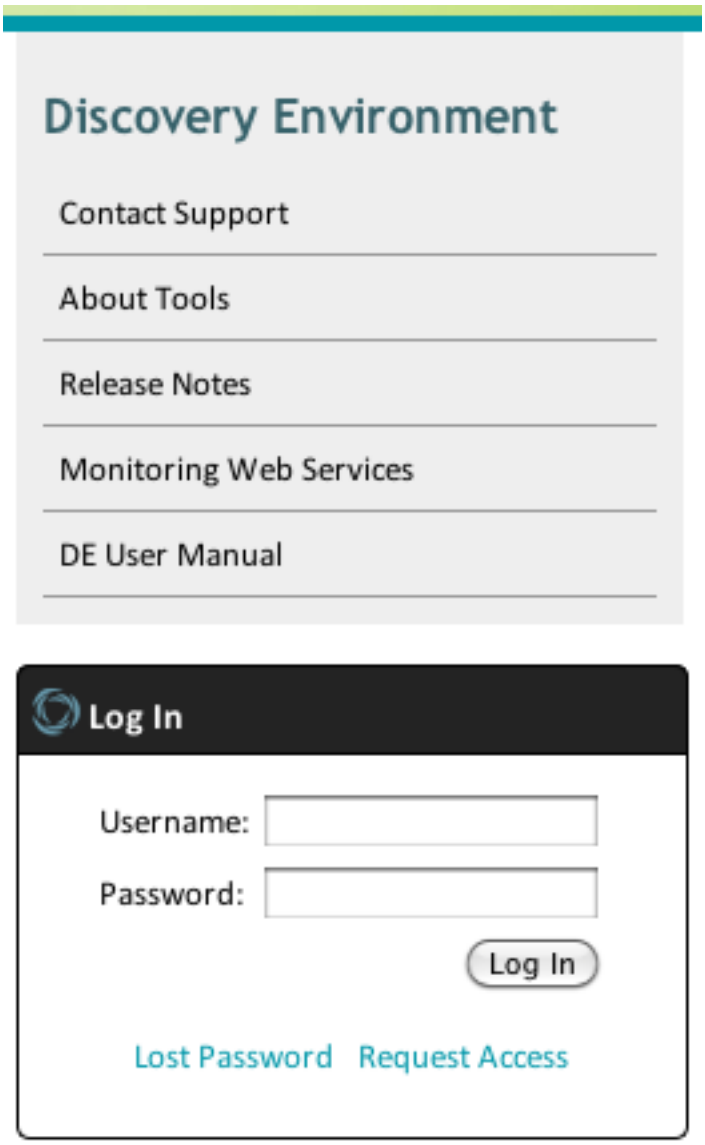
Fill out the form and click Submit.

When access is granted, you will receive a confirmation email that includes a link to create your password. You will not be able to log in until you create a password. This link can also be used to change your password.

Access the Log In Page

Access the Discovery Environment either by selecting the Discovery Environment link from the Tools window near the top of the iPlant homepage or by hovering your mouse cursor over the Discovery Environment Preview tab on the home page and clicking Log In to DE.

Login



The log in page contains a box on the left with some links for information related to the Discovery Environment, and the Log In box. To the right of this you will find a definition of a Discovery Environment.

Enter your username and password in the boxes provided on the left of the page. Click the Log In button to enter the environment.

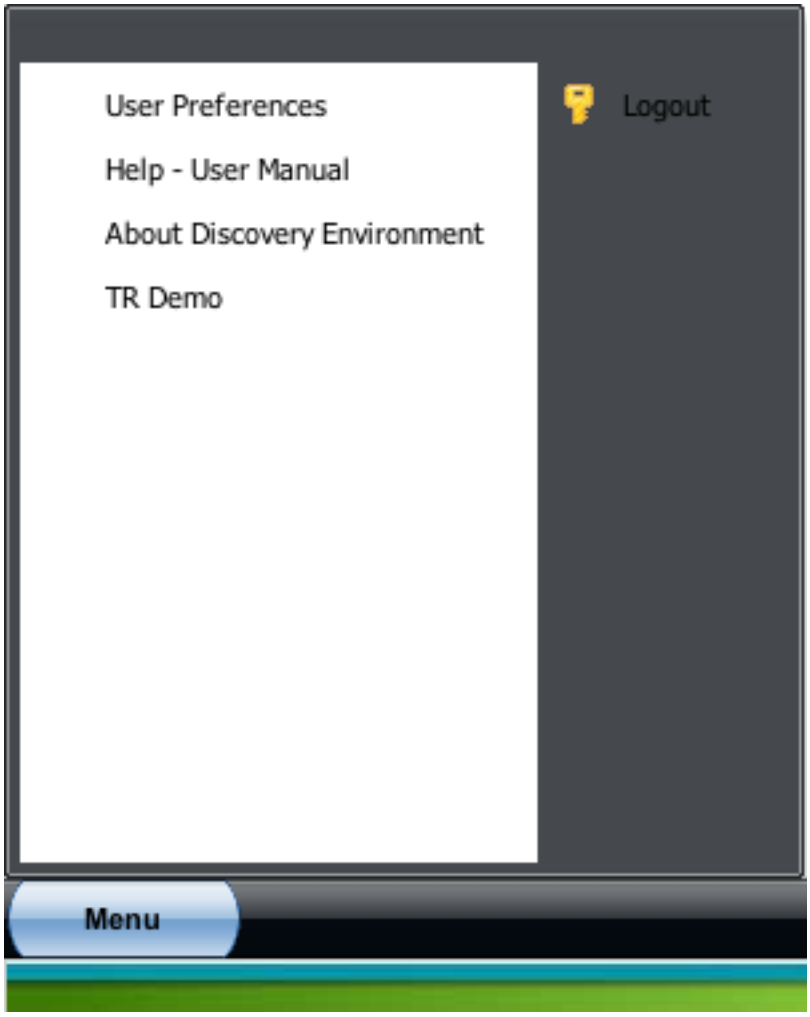
Click Lost Password if you need to reset your password. Click Request Access to access the same web form described earlier to request access to the Discovery Environment.

The Workspace

iPlant Collaborative™ Discovery Environment
Show all notifications
Manage Data
Perform Analyses
Menu
©2010 iPlant Collaborative
The iPlant Collaborative is funded by a grant from the National Science Foundation Plant Cyberinfrastructure Program (#DBI-0735191).

The Discovery Environment provides a consistent user interface and access to the high performance computing resources needed for specialized scientific analyses.

The Menu



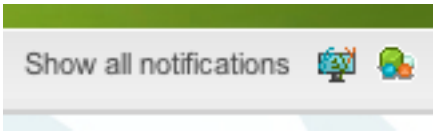
The Menu, available from the lower left corner of the Discovery Environment, is where you access some basic functions.

- User Preferences lets you update personal, institutional, and account information.
- Help - User Manual brings up the current version of this file.
- About Discovery Environment lists software details.
- TR Demo launches a demonstration preview version of our Tree Reconciliation tool.
- Logout will end your session.

Icons

Icons enable easy access to data and analyses.

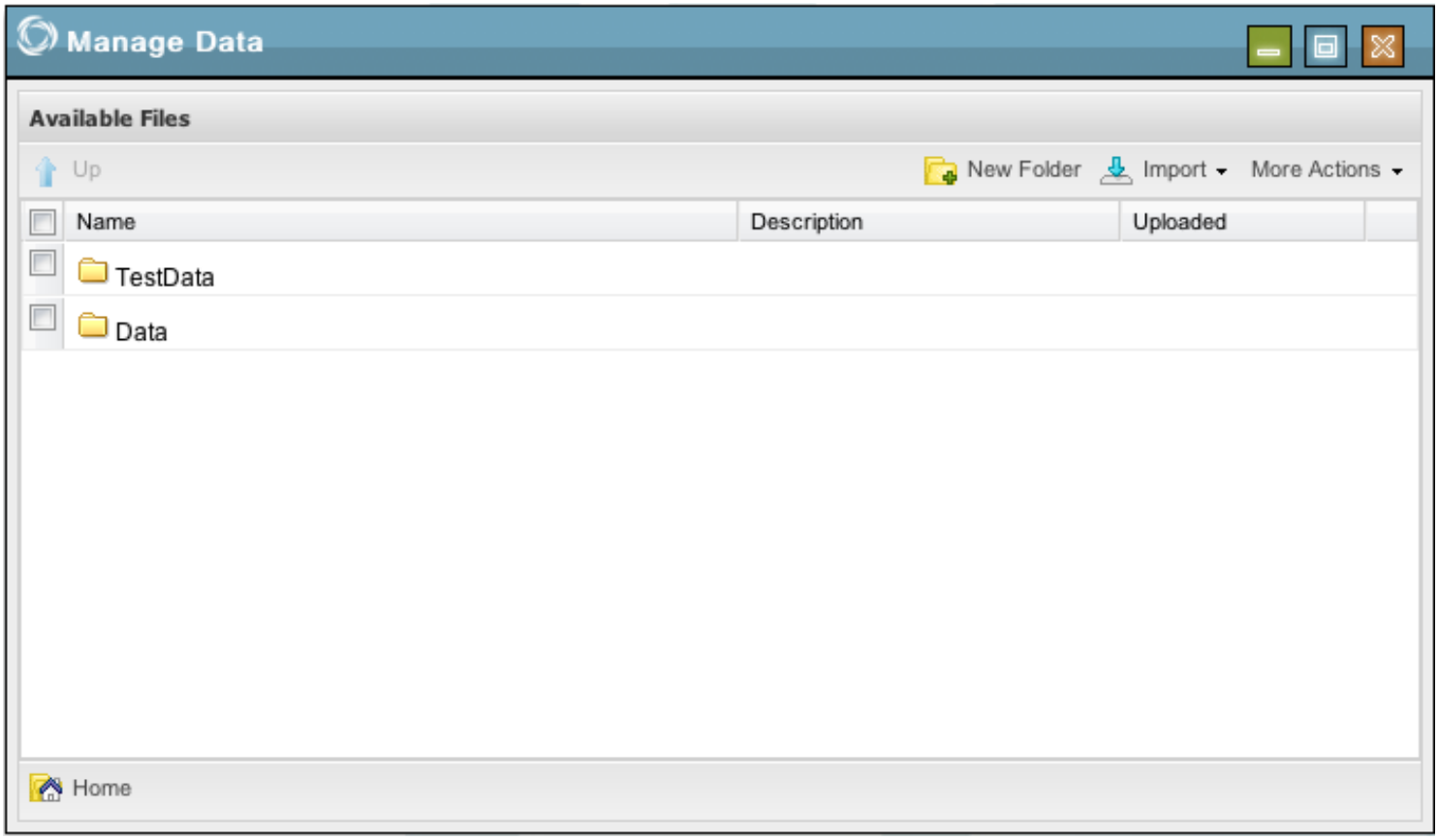
Notifications



Click Show all notifications to show messages from the system about status of data file imports and status updates for all analyses for your current session. The icons next to the text will sort those notifications by type, analysis, or data.

Manage Data

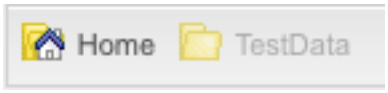
Introduction



Click the Manage Data icon to upload and manipulate data files.

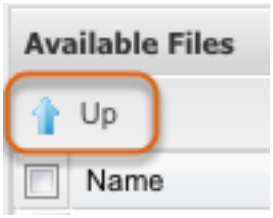
The window displays all files that you have uploaded or imported into the Discovery Environment, as well as some sample data provided to you by iPlant in the TestData folder.

Home Icon



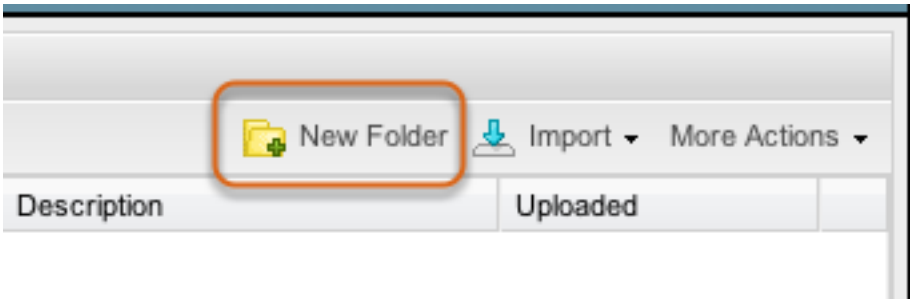
The Home icon at the lower left corner will always return you to the top level. When browsing folders, your current folder will appear next to this icon.

Up icon



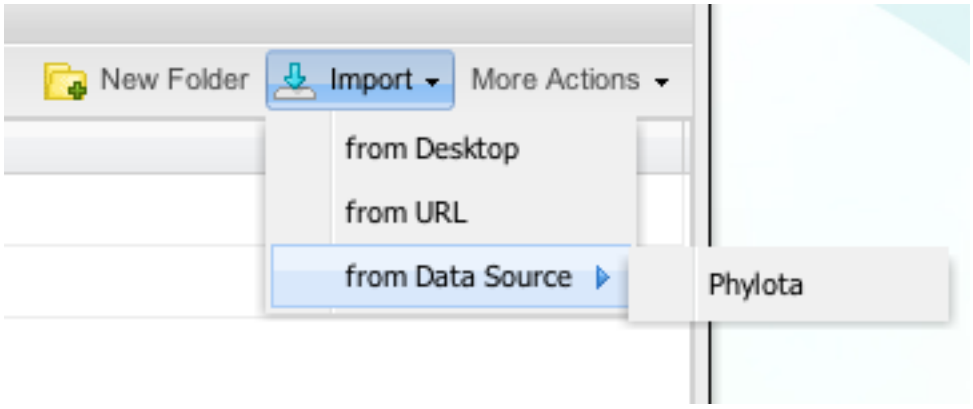
When browing within a folder, an Up icon will appear to the upper left of the list of files and folders. Click this to navigate one level above your current location.

Create a folder



Click New Folder to create a new folder in your current location.

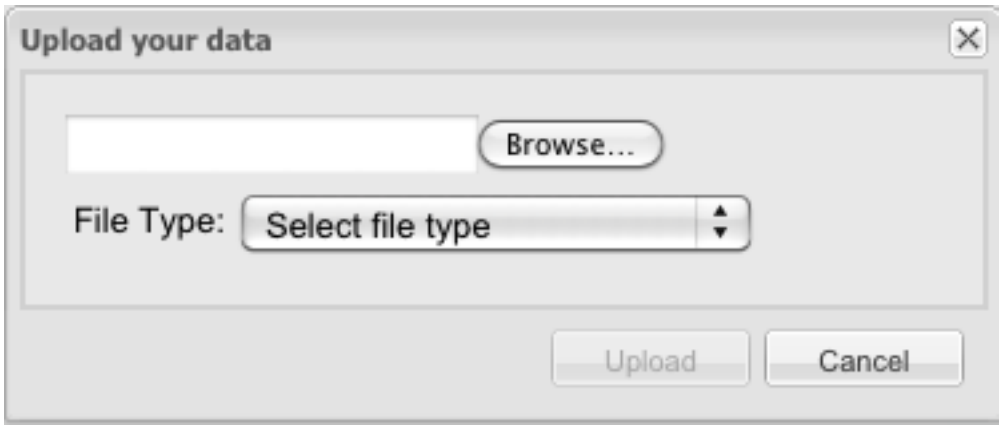
Import data



Import provides a drop-down menu from which you can upload data from your computer, import data from a URL, or import data from external repositories that have been enabled for direct access from the Discovery Environment.

Navigate to the folder into which you want to import data and click a menu option to import. Each method is described below.

Import from Desktop



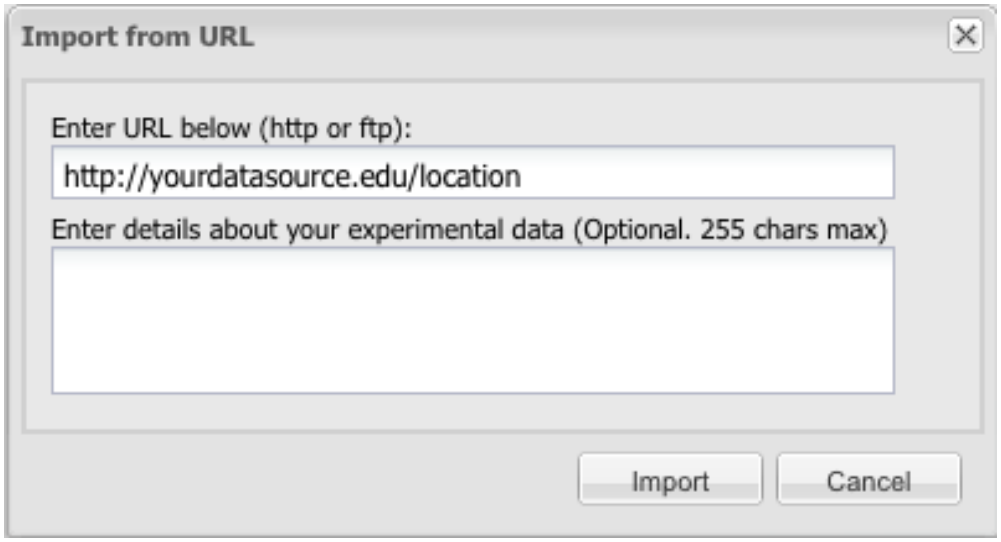
Click Browse to choose the file from your computer to import from your desktop. Select the appropriate file type from the drop-down list. Choices include Phylogenetic data, List of names for resolution, Sequence data, and Barcode file.

Click Upload.

File name restrictions for imported files:
File names must be unique and may be a maximum of 250 characters. All alphanumeric characters are permitted, along with these special characters: the dash (-), underscore (_), or period (.). Spaces are allowed, but are not permitted as the first, last, or only character.

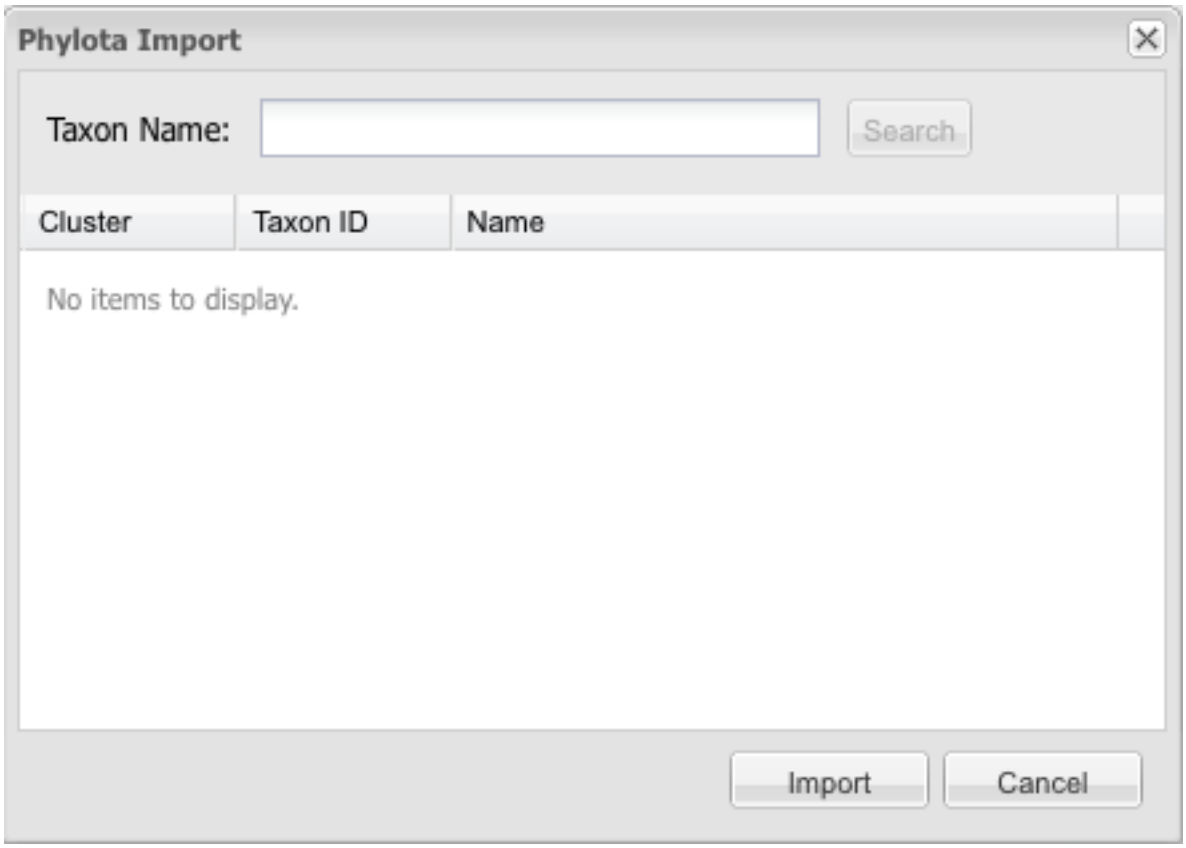
If a file is imported that has the same name as an existing file, the user is prompted that the file already exists and asked if he/she wants to overwrite. If yes, the file is imported as a new file.

Import from URL



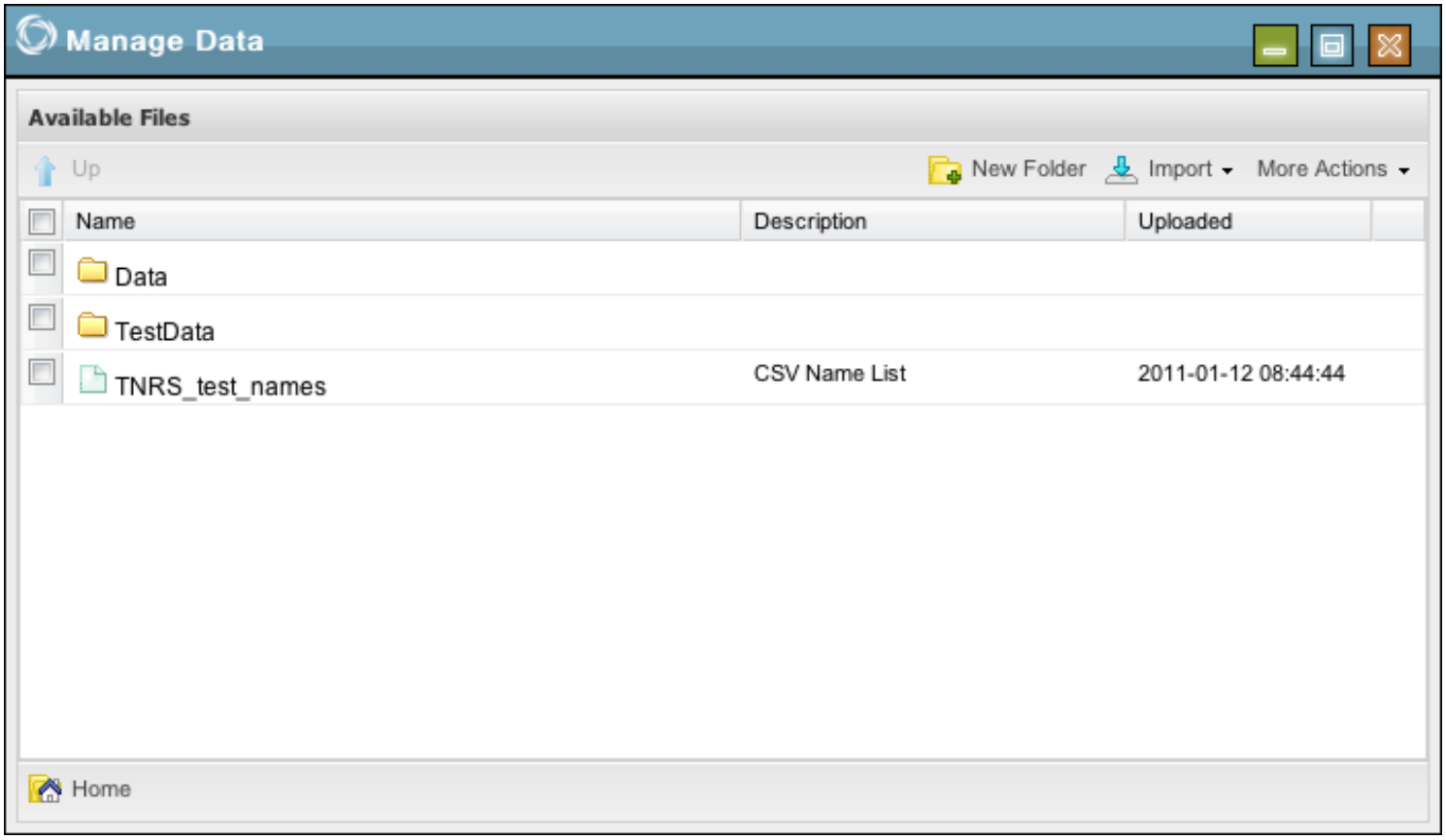
Enter the URL for the data file you wish to upload. Enter details about the data. Click Import.

Import from Data Source



You may currently import data from the Phylota database provided by the Sanderson lab at the University of Arizona. Enter the Taxon Name, click Search. Find the data you wish to import from the list and click Import.

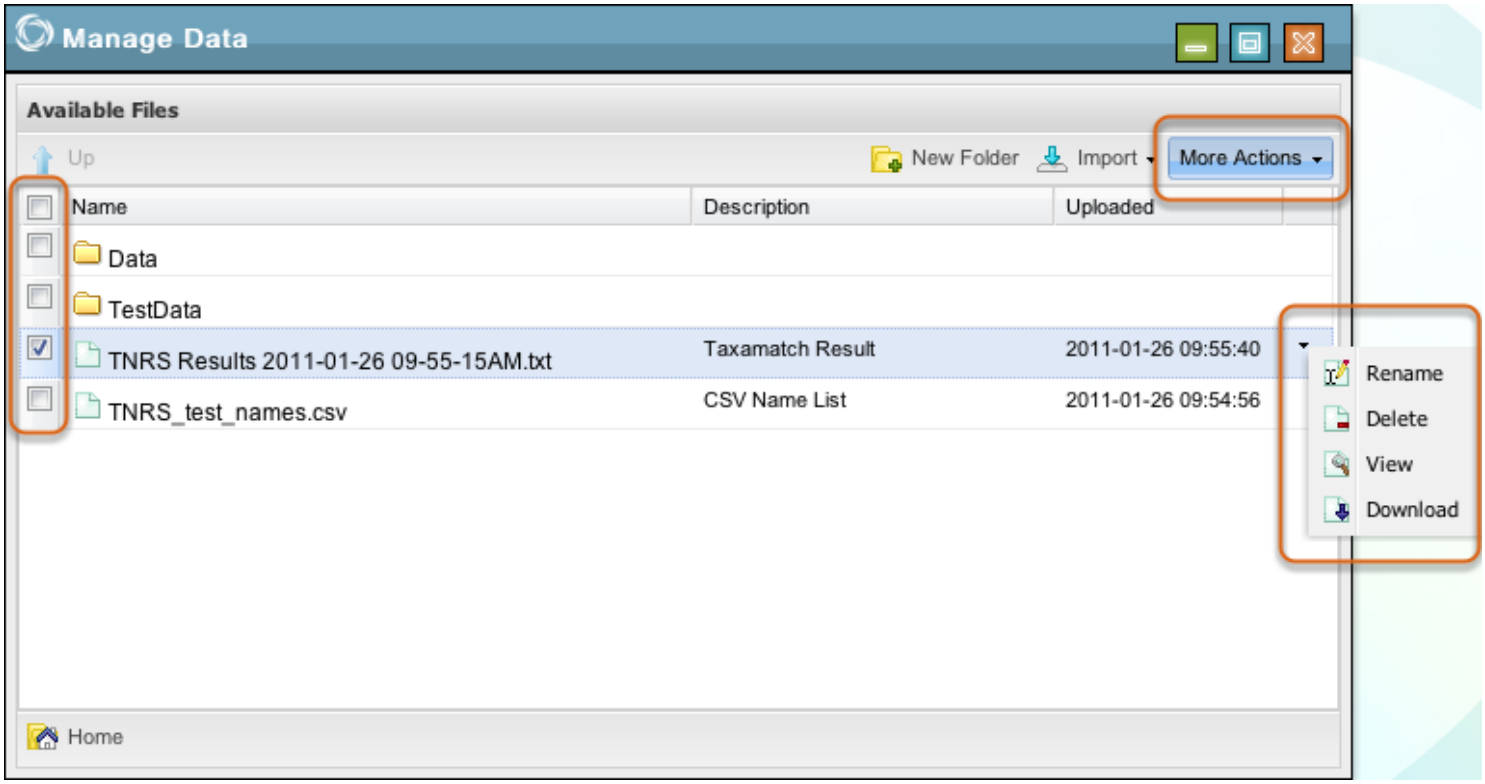
Confirm successful file import



Your file will appear in the list of available files in the folder you had open when you imported.

There are 3 verifications of successful import: a popup that flashes in the bottom right of the main screen, a notification in the Notifications list, and the file that appears in the selected folder in Manage Data.

More Actions

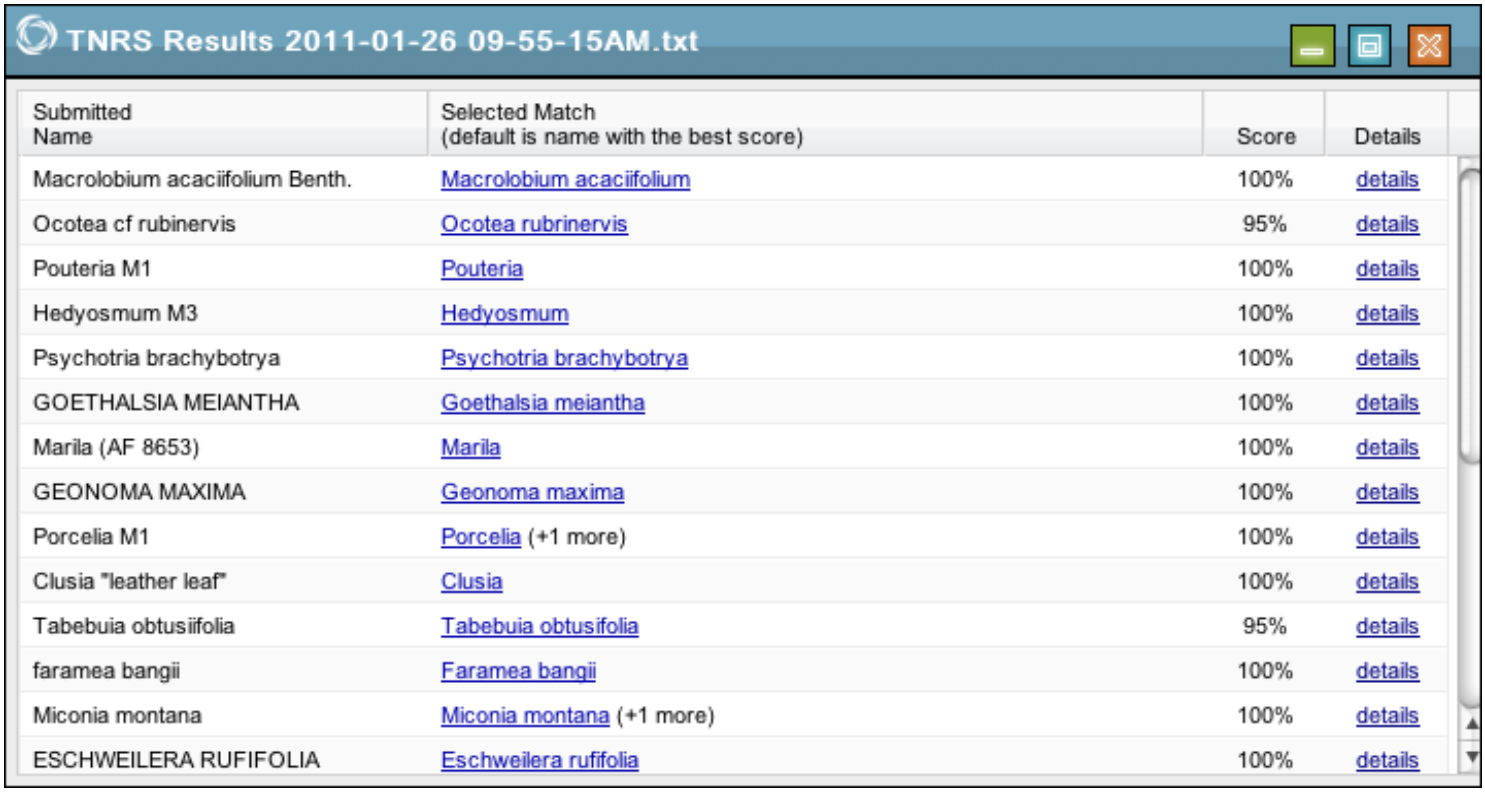


Mark the check box to the left of an item in this window to expose the drop-down menu shown at the right. Choose the appropriate entry to rename or delete files or folders, or to view or download file contents.

If you select one or more checkboxes, you may use the More Actions box or select an option in the menu to the right of any one of the selected items to perform the same tasks on single files or folders or some actions on multiple files or folders at the same time. Options are made available as follows:

- One file selected enables renaming, deleting, viewing or downloading the selected file.
- One folder selected enables renaming the selected folder or deleting it and all of its contents.
- More than one file selected enables deleting or viewing all selected files.
- More than one folder or a combination of files and folders selected enables deleting all the selected items.

View specific data



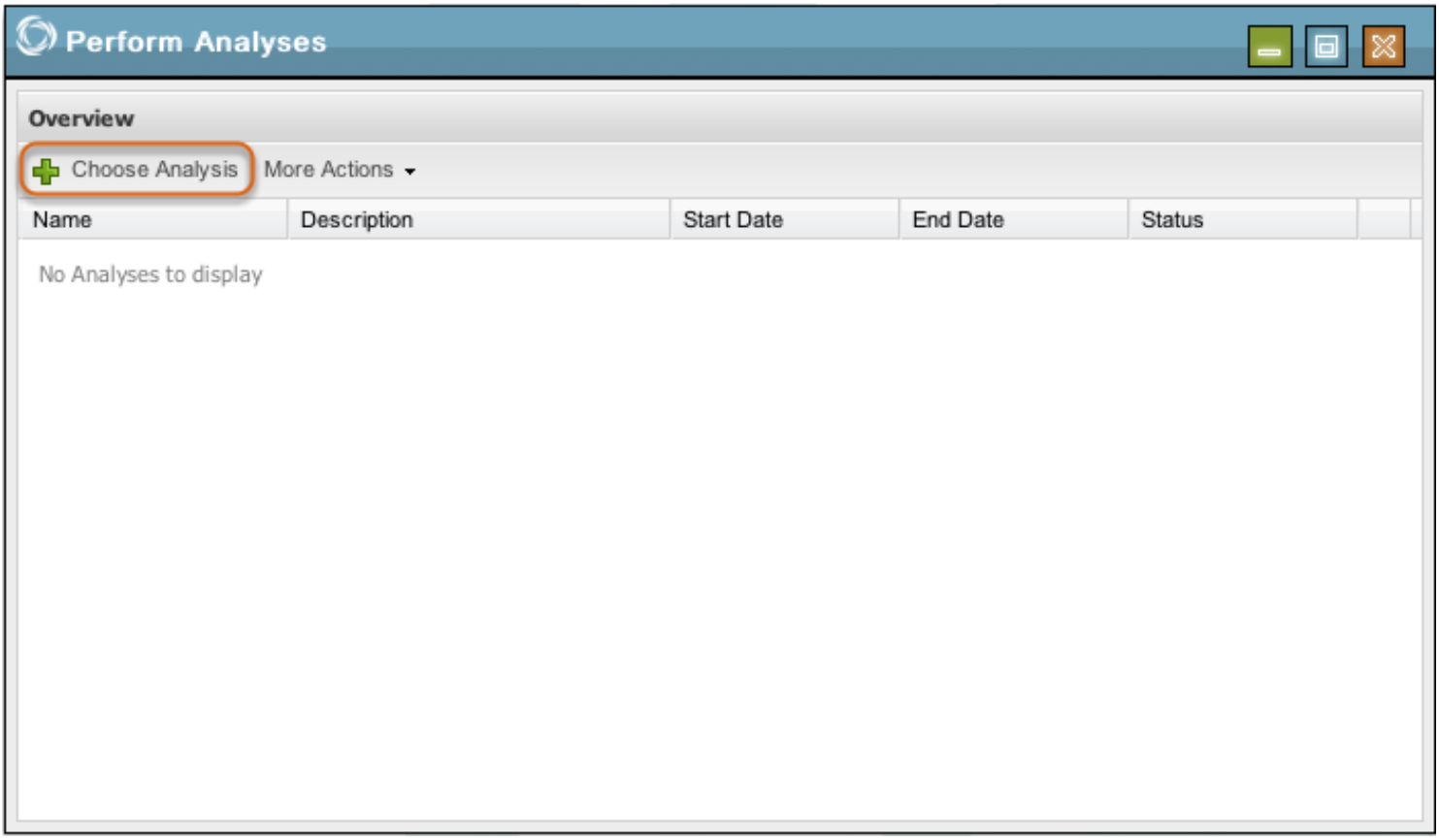
Submitted Name	Selected Match (default is name with the best score)	Score	Details
Macrolobium acaciifolium Benth.	Macrolobium acaciifolium	100%	details
Ocotea cf rubinervis	Ocotea rubrinervis	95%	details
Pouteria M1	Pouteria	100%	details
Hedyosmum M3	Hedyosmum	100%	details
Psychotria brachybotrya	Psychotria brachybotrya	100%	details
GOETHALSIA MEIANTHA	Goethalsia meiantha	100%	details
Marila (AF 8653)	Marila	100%	details
GEONOMA MAXIMA	Geonoma maxima	100%	details
Porcelia M1	Porcelia (+1 more)	100%	details
Clusia "leather leaf"	Clusia	100%	details
Tabebuia obtusiifolia	Tabebuia obtusifolia	95%	details
faramea bangii	Faramea bangii	100%	details
Miconia montana	Miconia montana (+1 more)	100%	details
ESCHWEILERA RUFIFOLIA	Eschweilera rufifolia	100%	details

Depending on the file selected, different tabs will appear in the new window.

For example, viewing a TNRS results file shows a list of names and matches with links. Viewing a .nex file will show Raw and Tree tabs. Viewing a .sam file will show Preview and Description tabs. Other file types display their contents in appropriate ways.

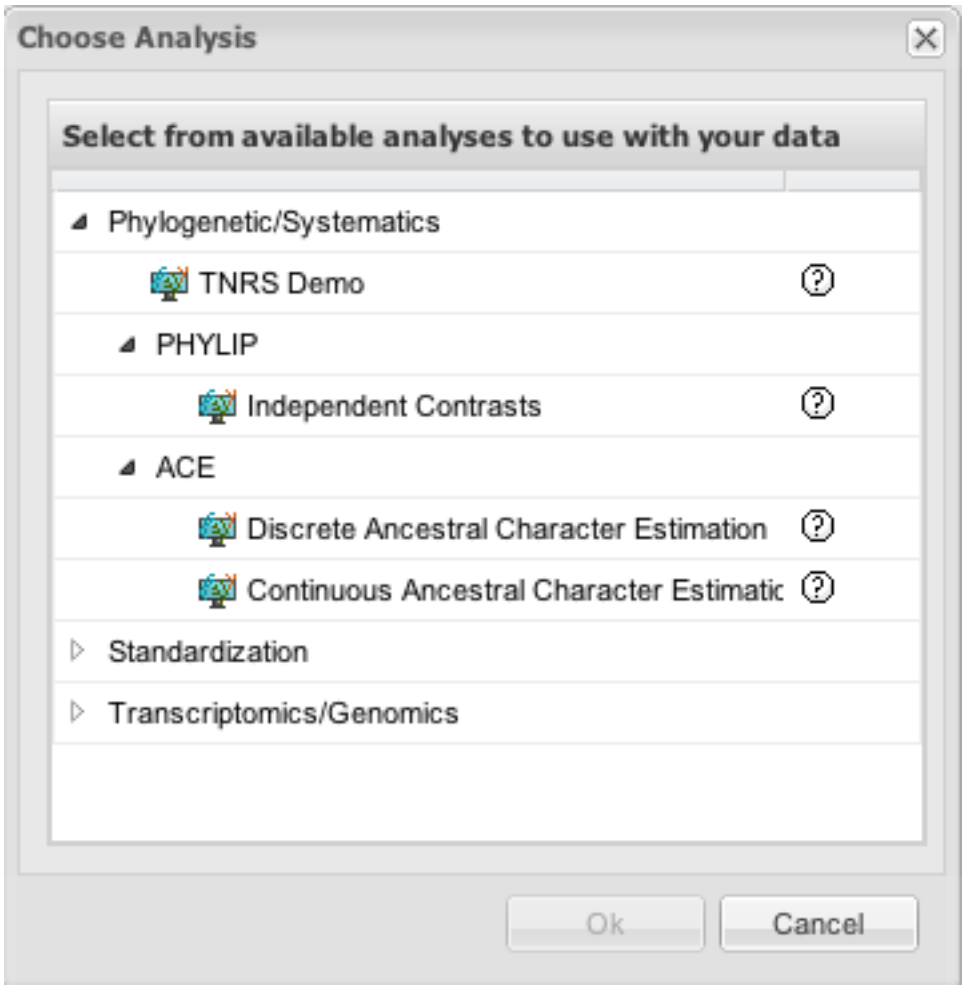
Perform Analyses

Analyses take implemented tools and enable them to be executed in the Discovery Environment. Click the Perform Analyses icon on the main page to start.



Perform Analyses is where you initiate analyses as well as view or delete completed analyses. Click Choose Analysis to initiate an analysis.

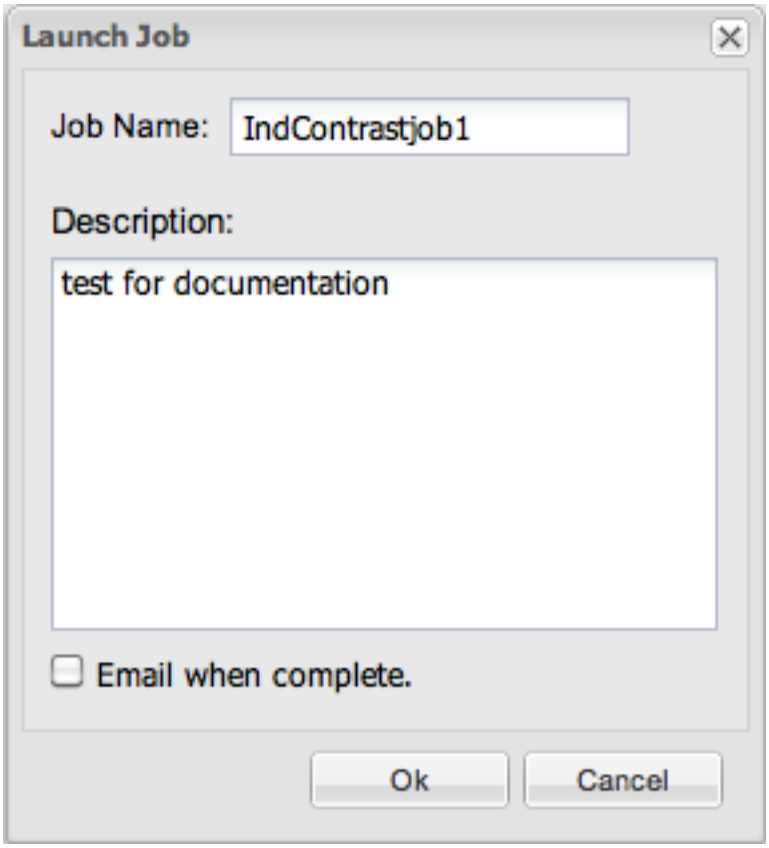
Choose Analysis



Analyses are categorized into logical groups to make specific tasks easier to find. Click the arrow next to a category to show what it contains. Select an analysis and click Ok to start.

When you have finished setting up your chosen analysis by following the steps it requires, click Launch Job.

Name Job

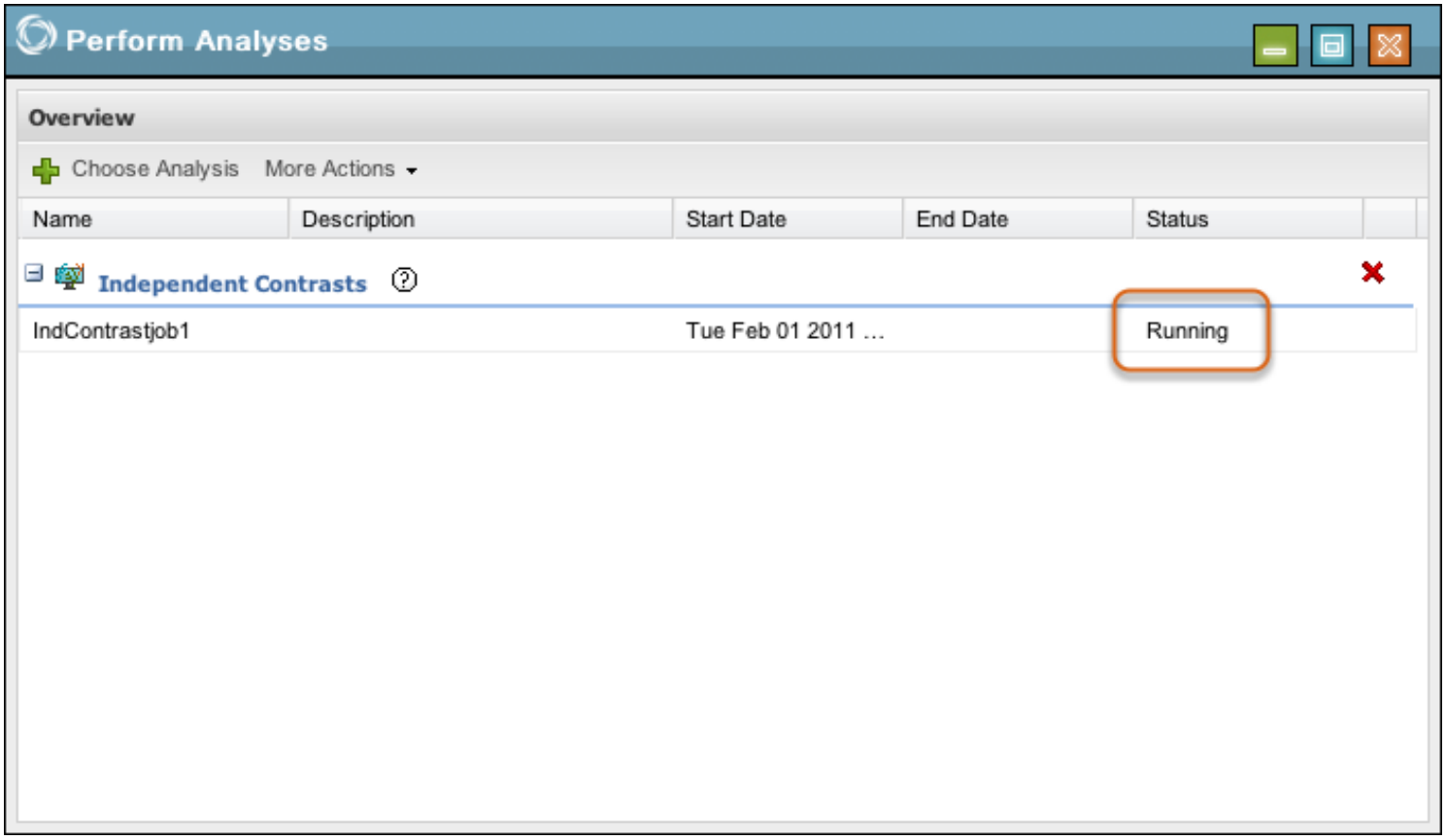


Enter a name for the job and write a description of it. The description is optional.

File name restrictions:
File names must be unique and may be a maximum of 250 characters. All alphanumeric characters are permitted, along with these special characters: the dash (-), underscore (_) or period (.). Spaces are allowed, but are not permitted as the first, last or only character.

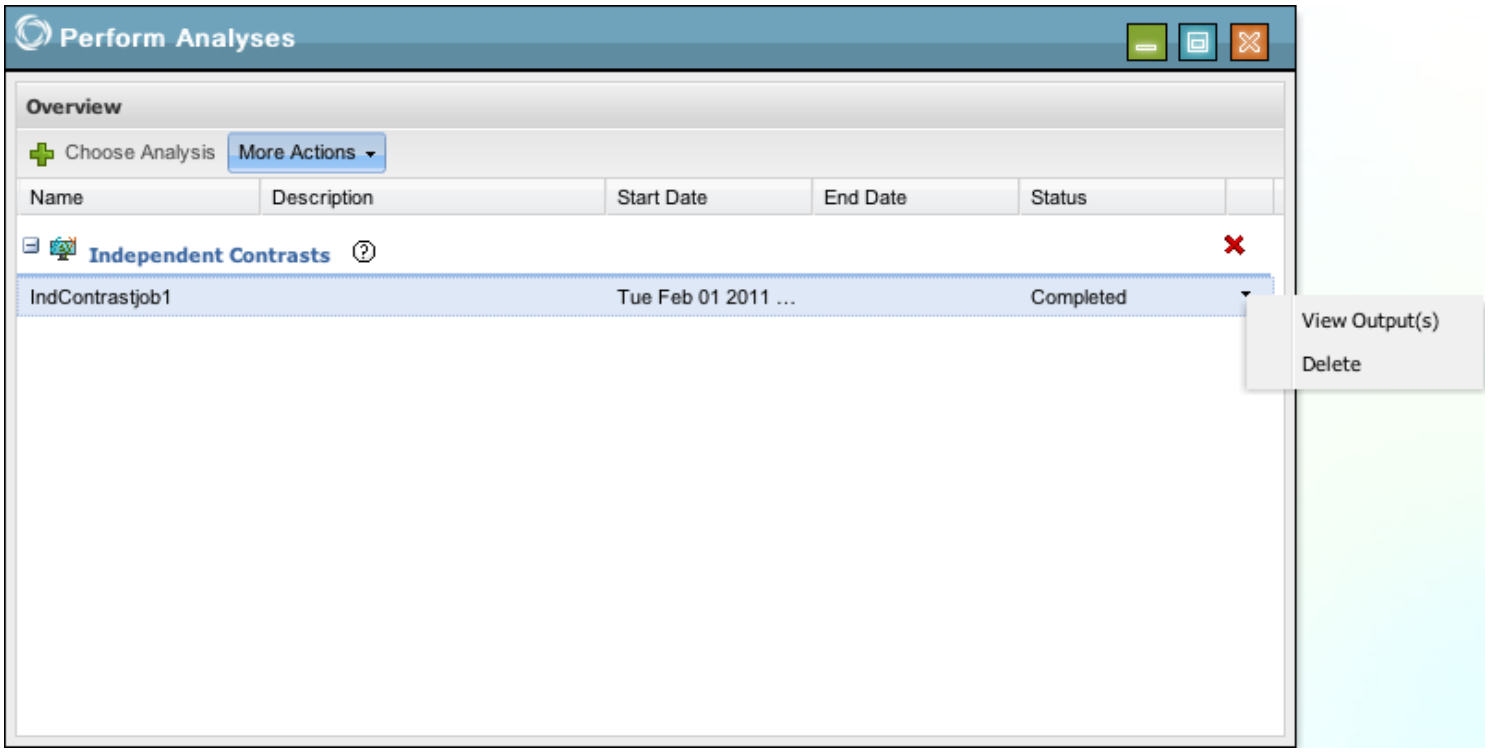
Click Ok to initiate your analysis.

View Analysis Status



When you run an analysis other than TNRS or TR, it will appear in Perform Analyses. The Status will update as the analysis is completed.

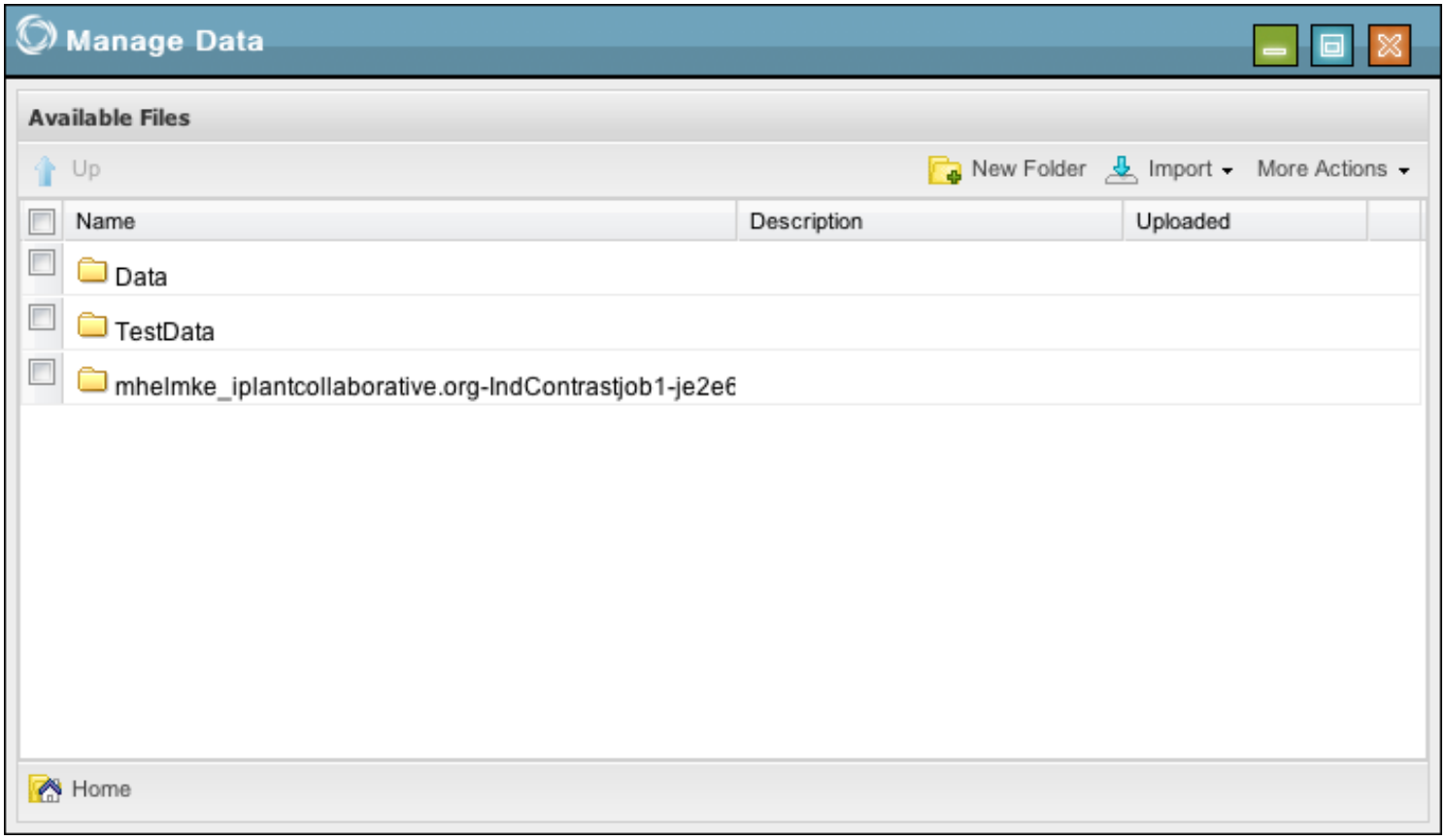
View Analysis Output(s)



After a completed run of an analysis, you can view the results. Select the analysis and then select View Output(s) from the drop-down menu at the right. You can also find View Output(s) under More Actions.

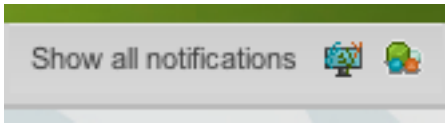
To delete a completed analysis, select Delete from the drop-down menu at the right or from More Actions.

View Output(s) (alternate)

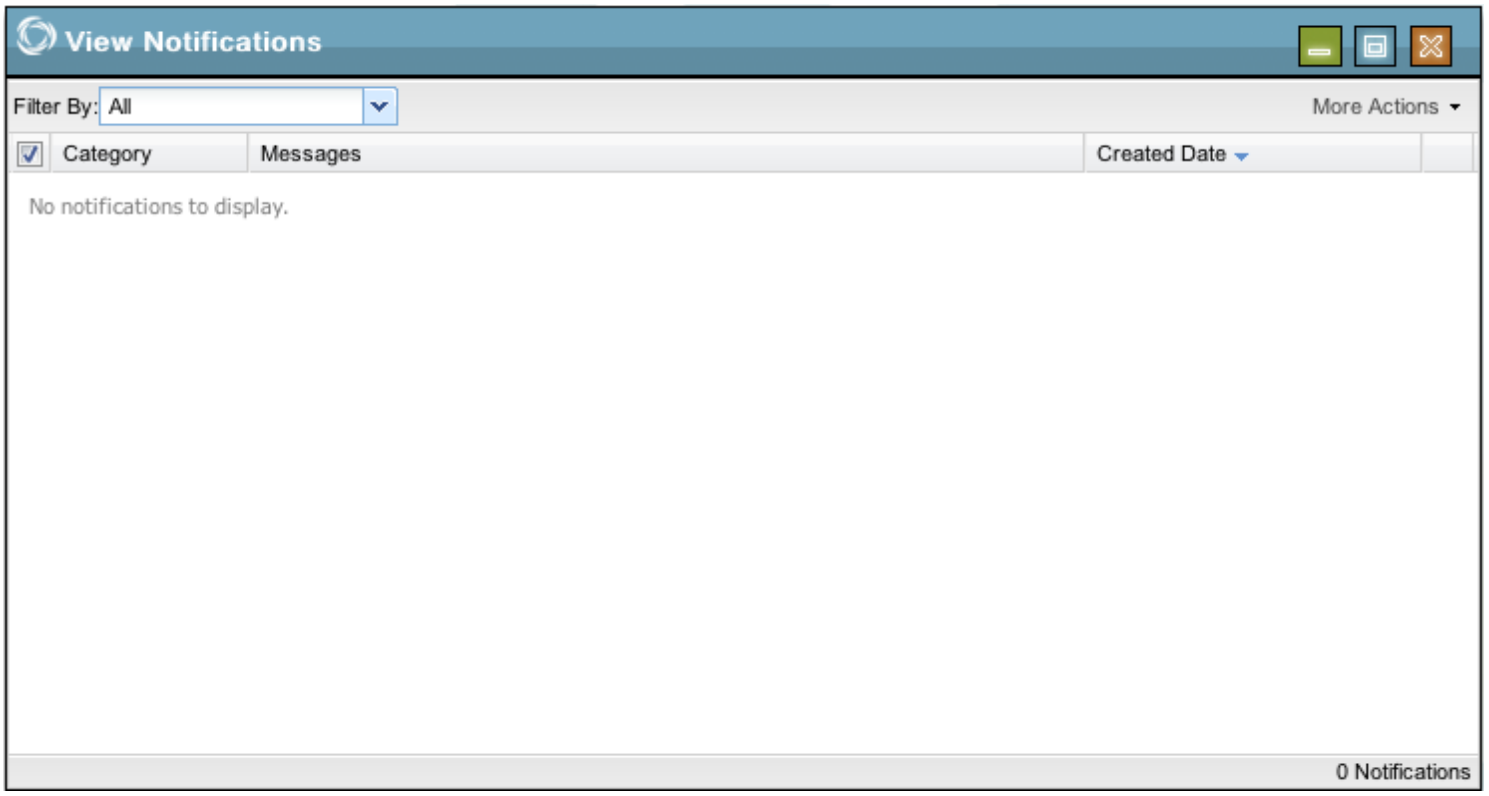


Analysis output(s) are automatically placed in a folder in Manage Data and may be viewed from there at any time after a completed run.

Viewing and Deleting Notifications

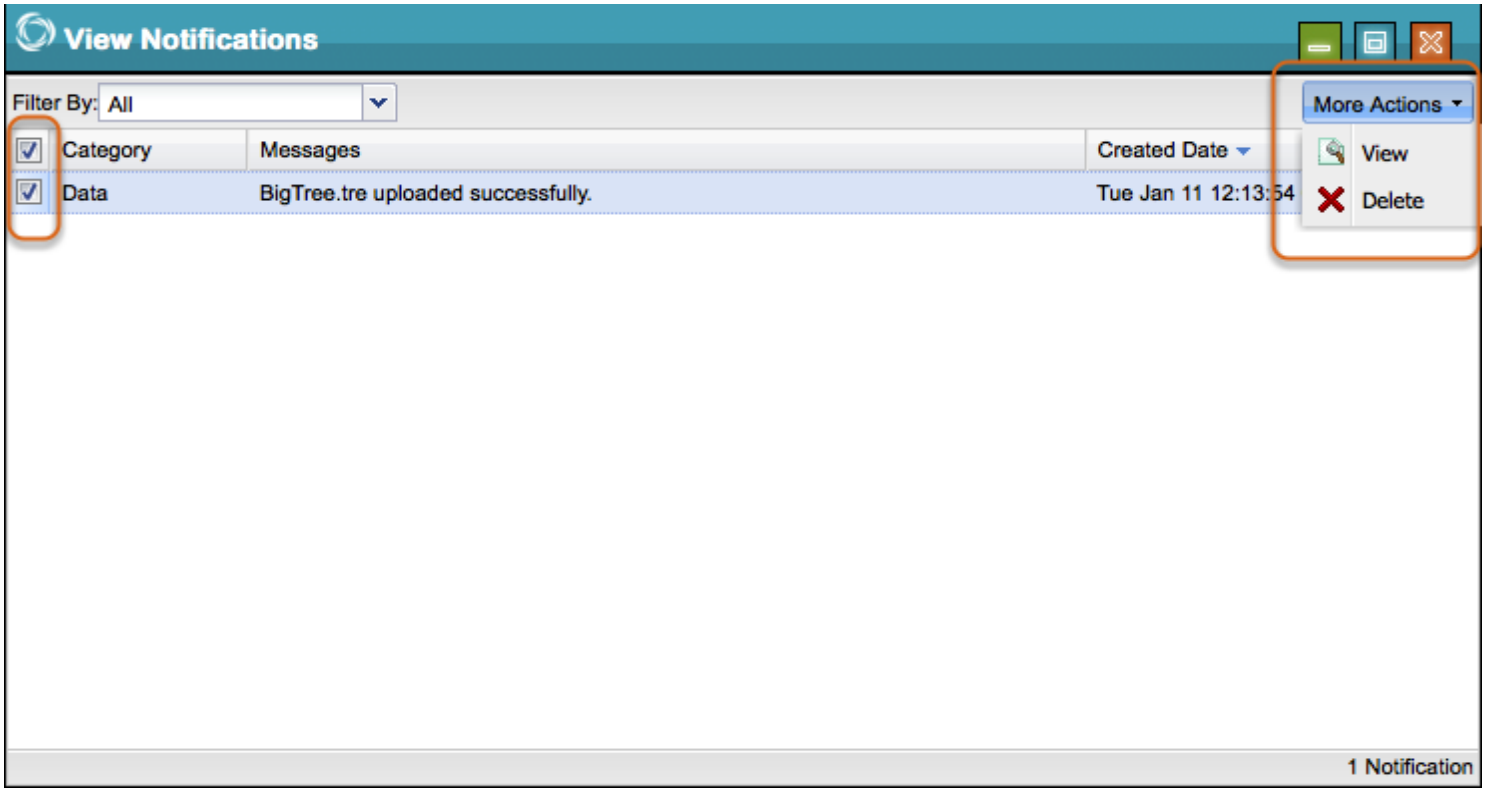


Click Show all notifications near the top right corner of the Discovery Environment screen to show messages from the system. The icons next to the text will sort those notifications by type, analysis, or data.



Notifications are shown in View Notifications and may be filtered by type using the Filter By drop-down menu.

More Actions



Use the checkboxes to select notifications.

Notifications that include other data, such as successful data imports and analysis results, may be viewed or deleted from the More Actions drop-down.

Notifications that merely inform, such as delete success notices that only appear as popups in the main window, are temporary and require no further action.

Analyses

Ancestral Character Estimation (ACE) Overview

An ancestral character is a biological trait that is present in a group of related organisms and is thus inferred to have been present in the most recent common ancestor of these organisms. Traits of interest, for example fruit size or the presence of parasite resistance, can therefore be traced back in time along a known phylogeney.

Estimating ancestral character values is a phylogenetic analysis that can be used to test evolutionary hypotheses like the temporal sequence of evolutionary events or the appearance of adaptive traits. Because ancestral characters values are not observed, it is more rational to consider them as parameters in a model where the character values of recent species are the observed values.

It is possible to perform both continuous and discrete ancestral character estimations in the Discovery Environment. Both use a software package called ape, which is based on R, to perform estimation based on a fully resolved phylogeny.

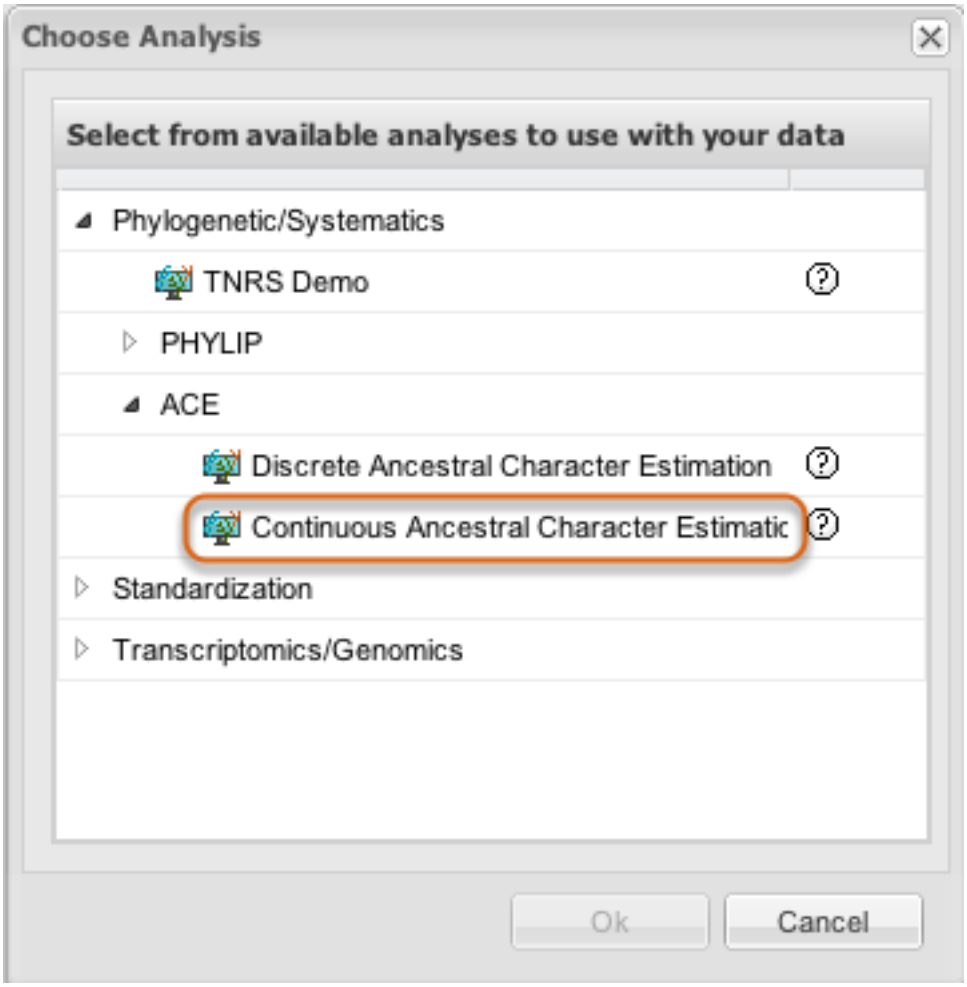
Continuous ancestral character estimation (CACE) assumes that traits evolve according to a Brownian motion process. Under this model, the expected difference between two taxa can be computed as a function of the time separating the taxa from their most recent common ancestor, which is obtained from the phylogenetic tree. Maximum Likelihood is then used to obtain the ancestors' trait values, which minimizes the sum of squared changes along the branches. The output is a table of ancestral trait values and the corresponding 95% confidence intervals. These value estimates can be plotted on the phylogenetic tree using a color gradient. Additionally, the function outputs an estimate of the Brownian motion parameter σ2 and the log likelihood of the model.

Discrete ancestral character estimation (DACE) describes evolutionary trait changes using a continuous-time Markov model. In this model the probability of change from one state to another depends only on the transition rate and the evolutionary time, which is obtained by the phylogeny. Maximum Likelihood is then used to estimate the transition rates and the proportional likelihoods of the ancestor's states. The output is a table of proportional likelihood for all possible states at the internal nodes. These value estimates can be plotted on the phylogenetic tree using pie charts to represent the likelihoods. Additionally, the function outputs an estimate of the transition rate with its associated uncertainty and the log likelihood of the model.

More details about ape can be found at:
http://cran.r-project.org/web/packages/ape/index.html
http://ape.mpl.ird.fr/

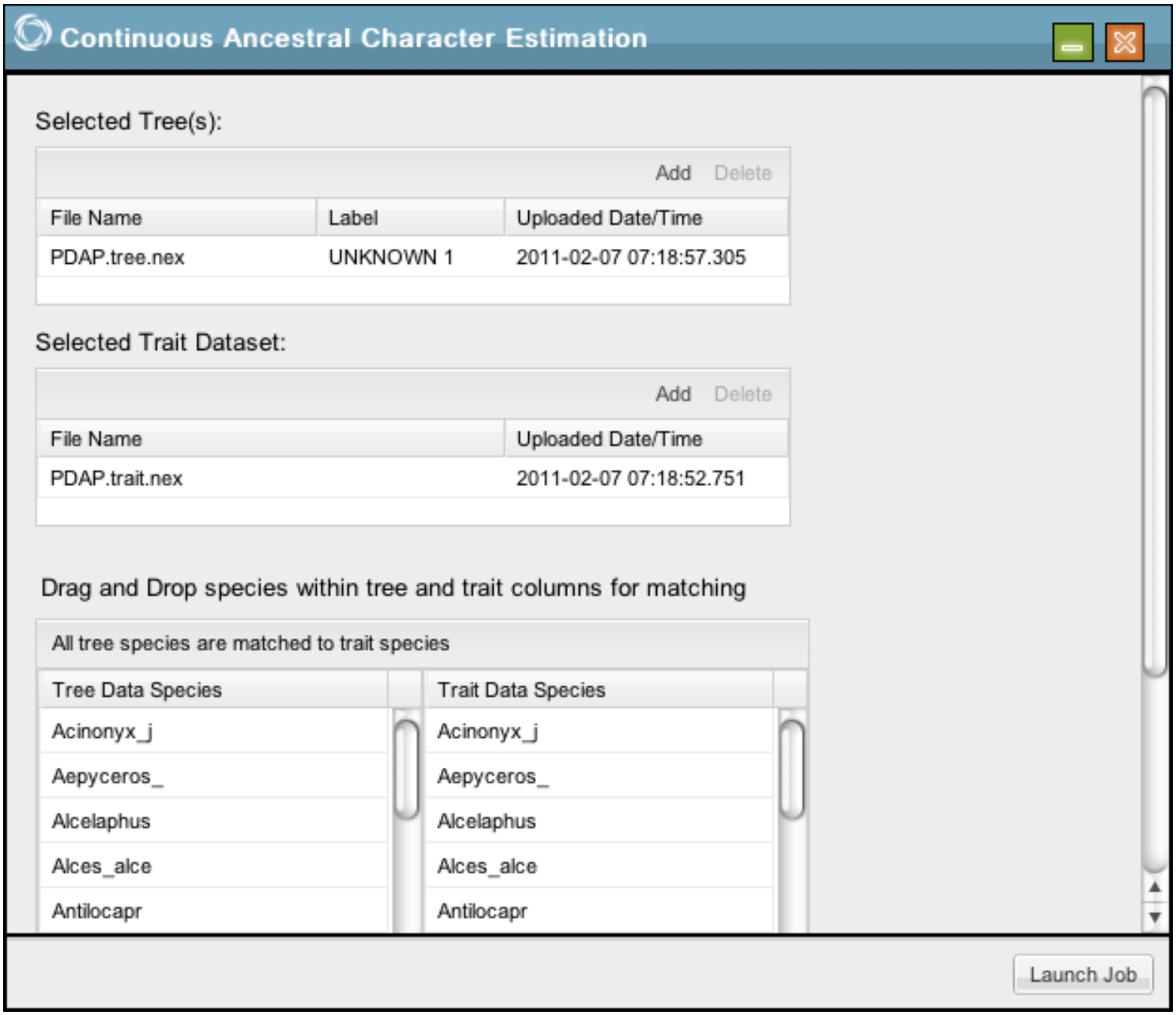
Continuous Ancestral Character Estimation (CACE)

An overview of Ancestral Character Estimation is available.



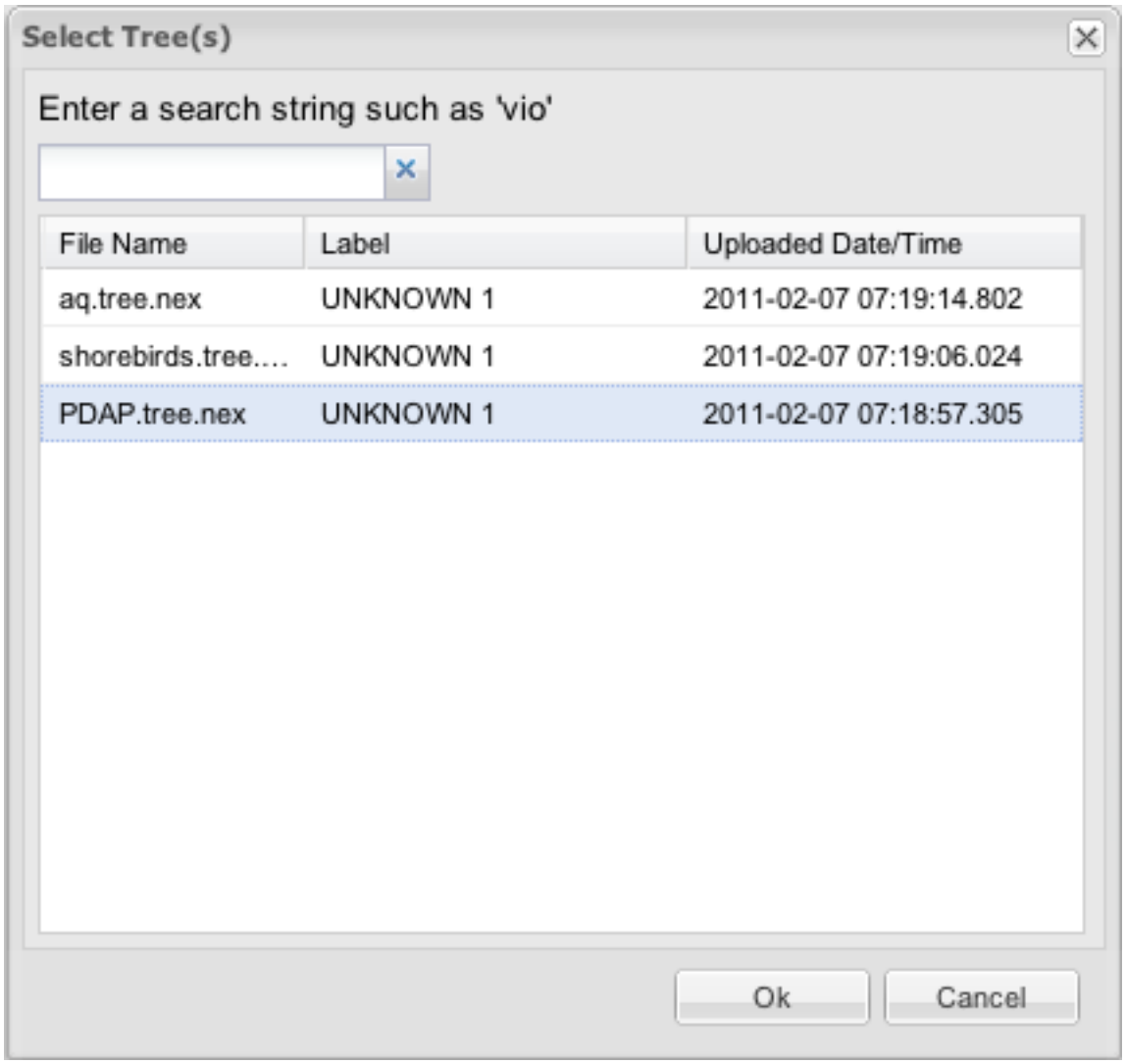
Select Continuous Ancestral Character Estimation (CACE) from within Perform Analyses as described in that section. Click Ok.

Select data

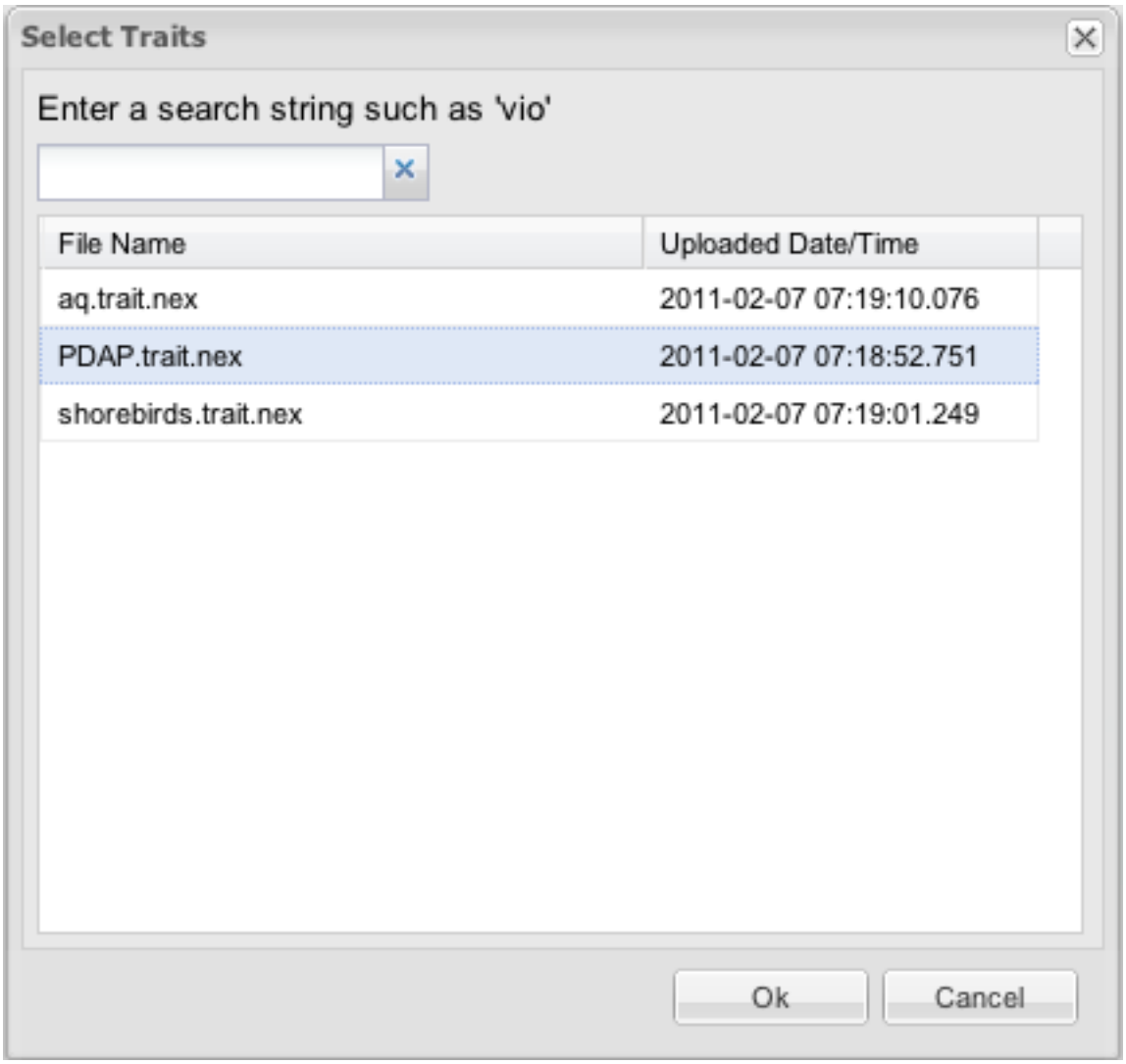


Data needs to be uploaded to the Discovery Environment in advance. Click Add in Selected Tree(s) and Selected Trait Dataset to choose appropriate tree and trait files from the boxes shown next.

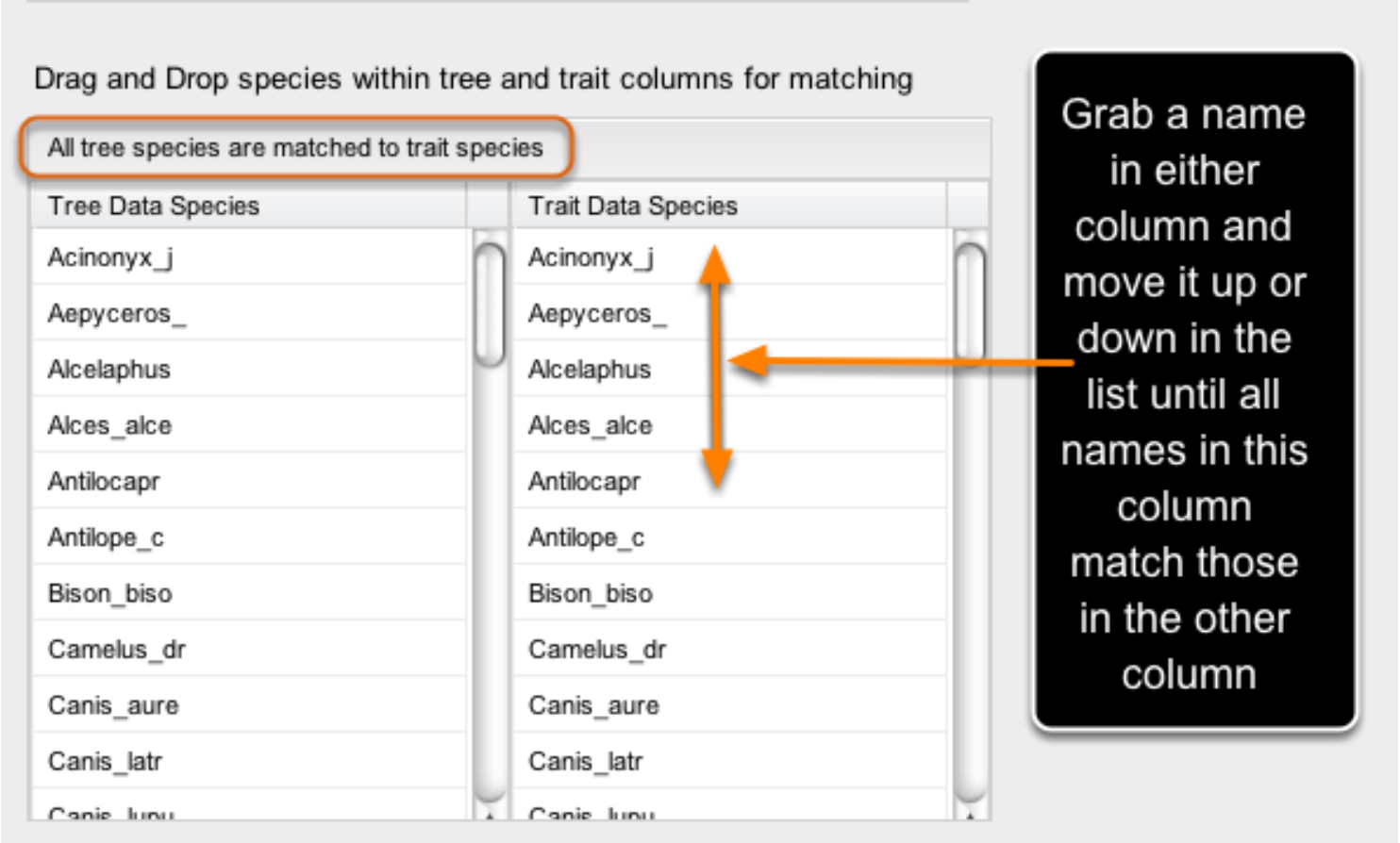
Select Tree or Trees



Select Traits



Match Data

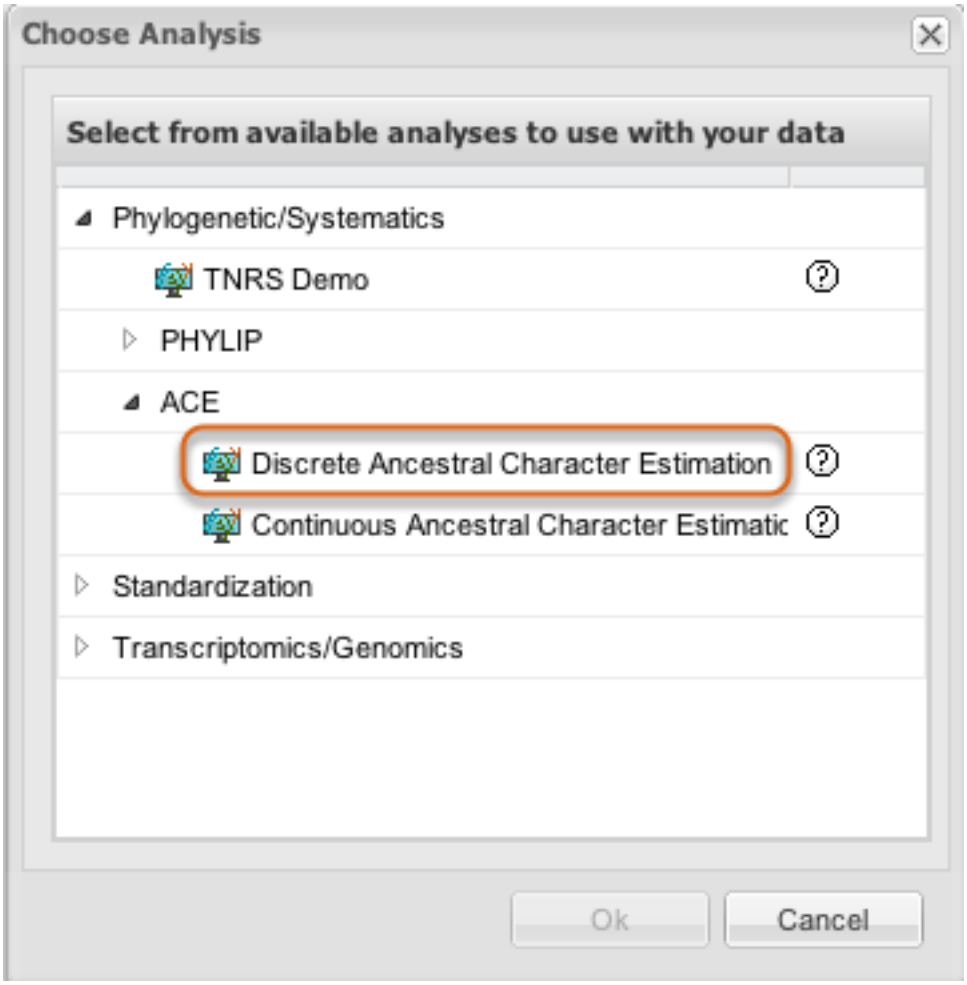

Drag and Drop species within tree and trait columns for matching
All tree species are matched to trait species
Tree Data Species
Trait Data Species
Acinonyx_j
Aepyceros_
Alcelaphus
Alces_alce
Antilocapr
Antilope_c
Bison_biso
Camelus_dr
Canis_aure
Canis_latr
Grab a name in either column and move it up or down in the list until all names in this column match those in the other column

Hold the left mouse button to drag and swap to move species data up and down until all tree species and trait species are matched. When the text above the table shows All tree species are matched to trait species, click Launch Job.

Enter a name and description for the job and click Ok. See Perform Analyses for information about monitoring the process and where to find your results.

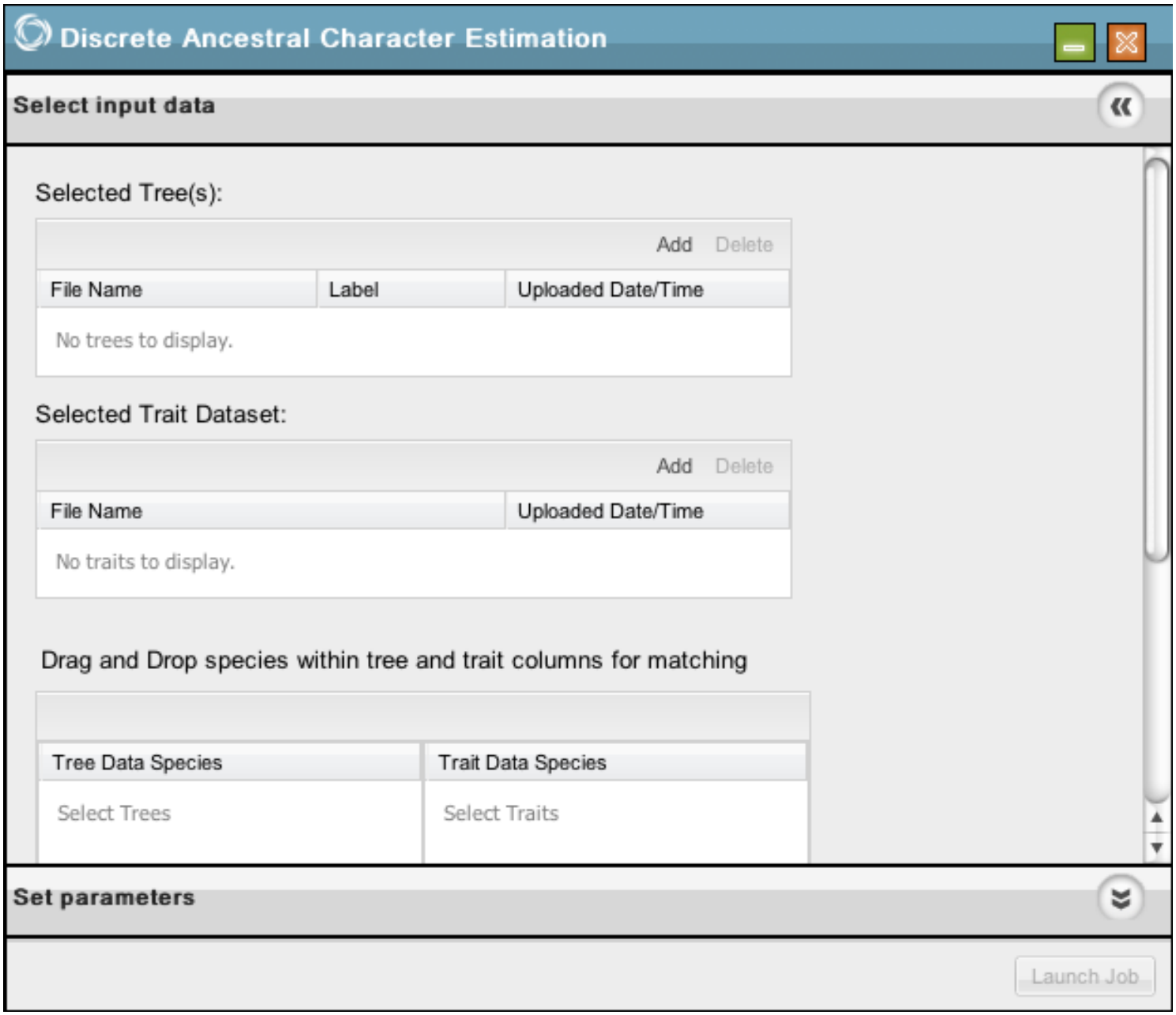
Discrete Ancestral Character Estimation (DACE)

An overview of Ancestral Character Estimation is available.



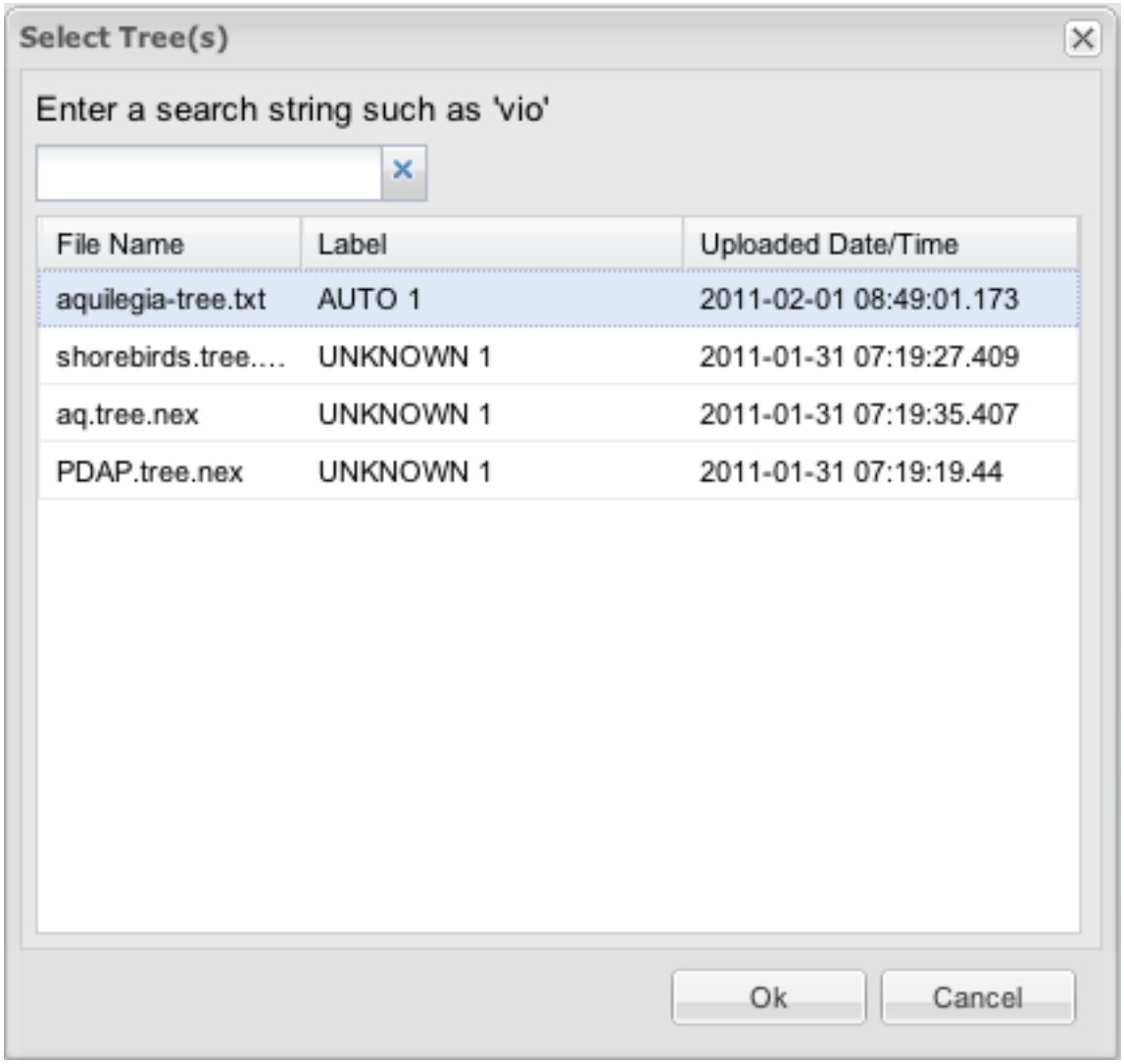
Select Discrete Ancestral Character Estimation from within Perform Analyses as described in that section. Click Ok.

Select data

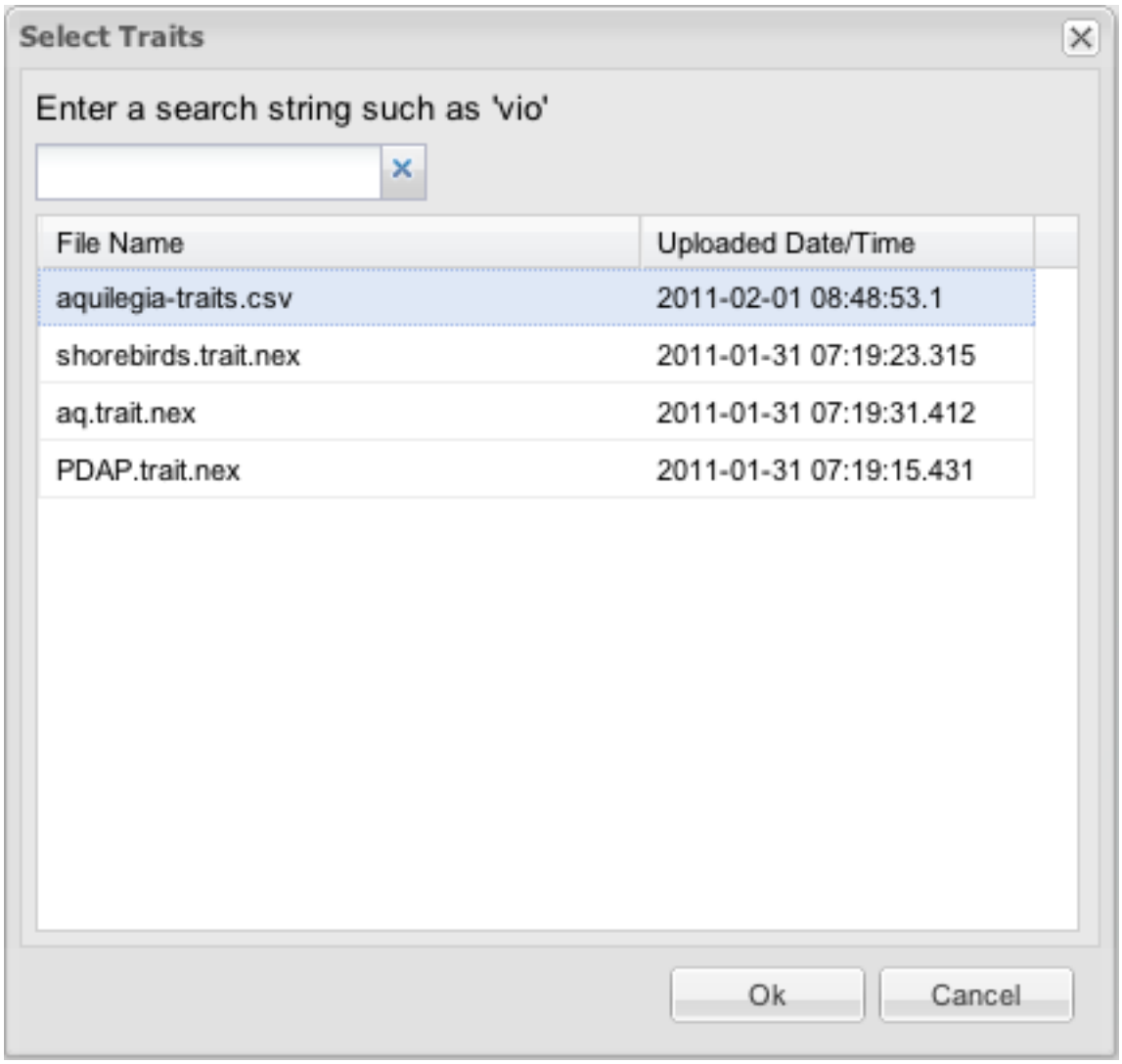


Data needs to be uploaded to the Discovery Environment in advance. Click Add in Selected Tree(s) and Selected Trait Dataset to choose appropriate tree and trait files from the boxes shown next.

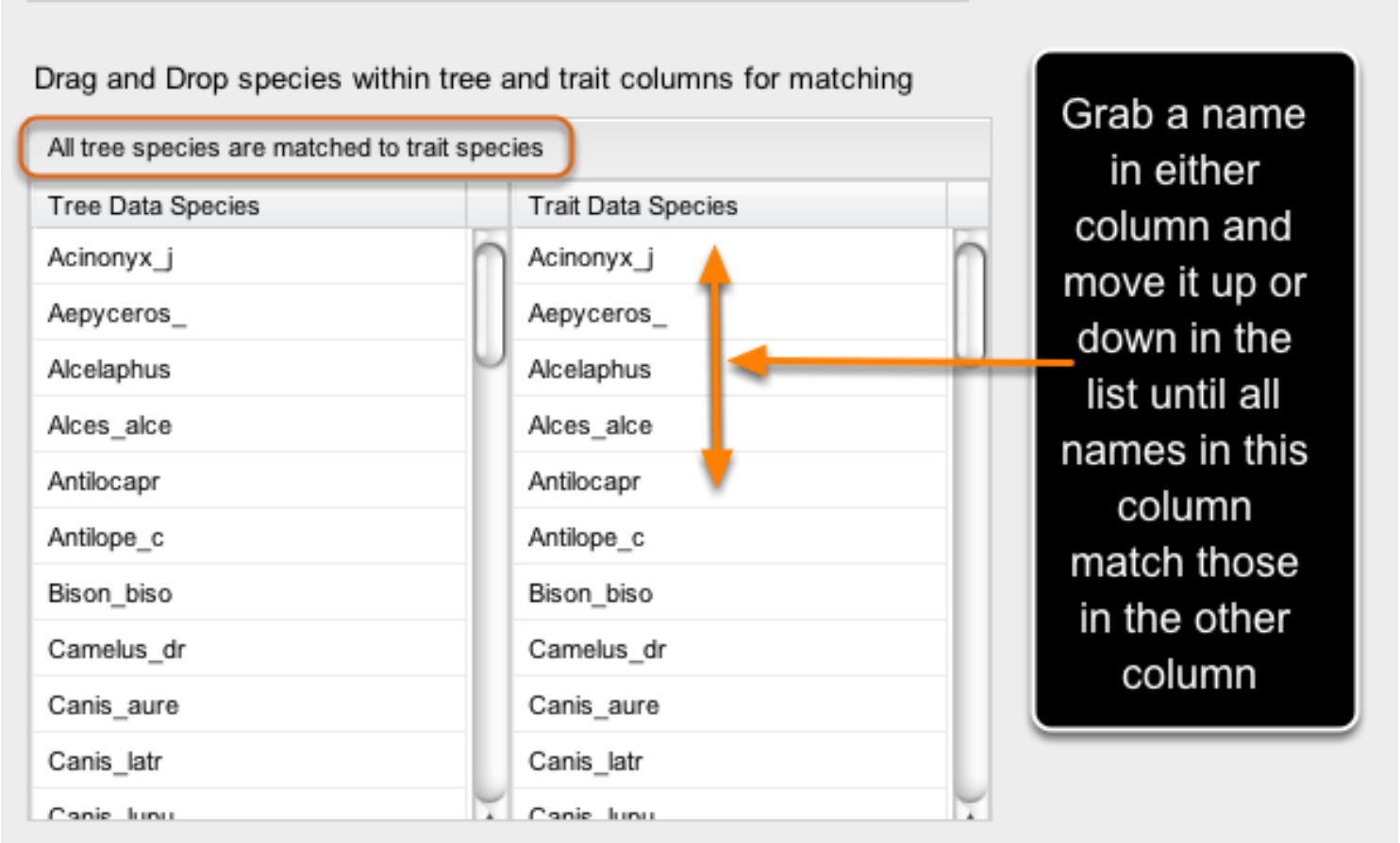
Select Tree or Trees



Select Traits

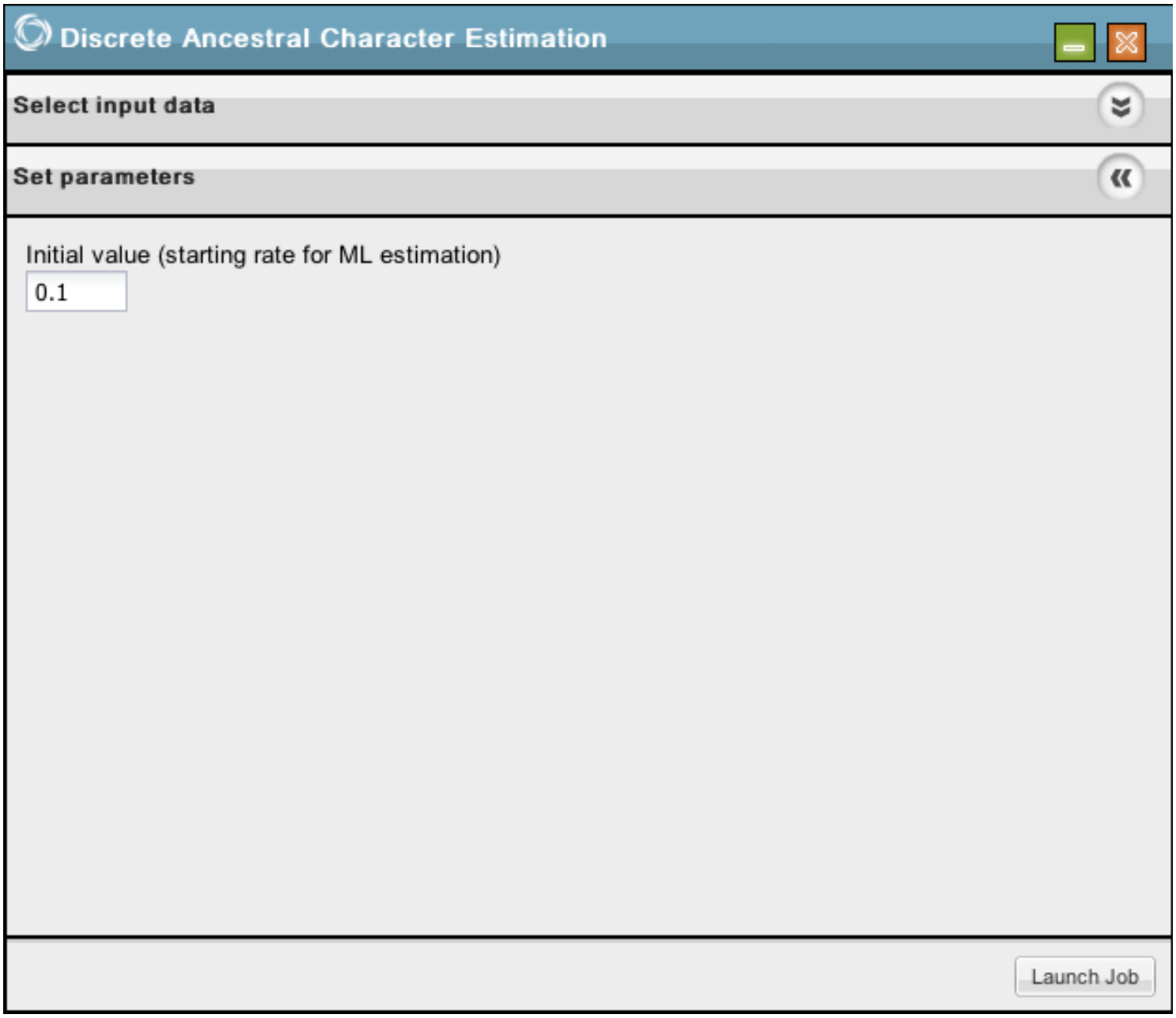


Match Data


Drag and Drop species within tree and trait columns for matching
All tree species are matched to trait species
Tree Data Species
Trait Data Species
Acinonyx_j
Aepyceros_
Alcelaphus
Alces_alce
Antilocapr
Antilope_c
Bison_biso
Camelus_dr
Canis_aure
Canis_latr
Grab a name in either column and move it up or down in the list until all names in this column match those in the other column

Hold the left mouse button to drag and swap to move species data up and down until all tree species and trait species are matched. When the text above the table shows All tree species are matched to trait species, click Select output details.

Set parameters



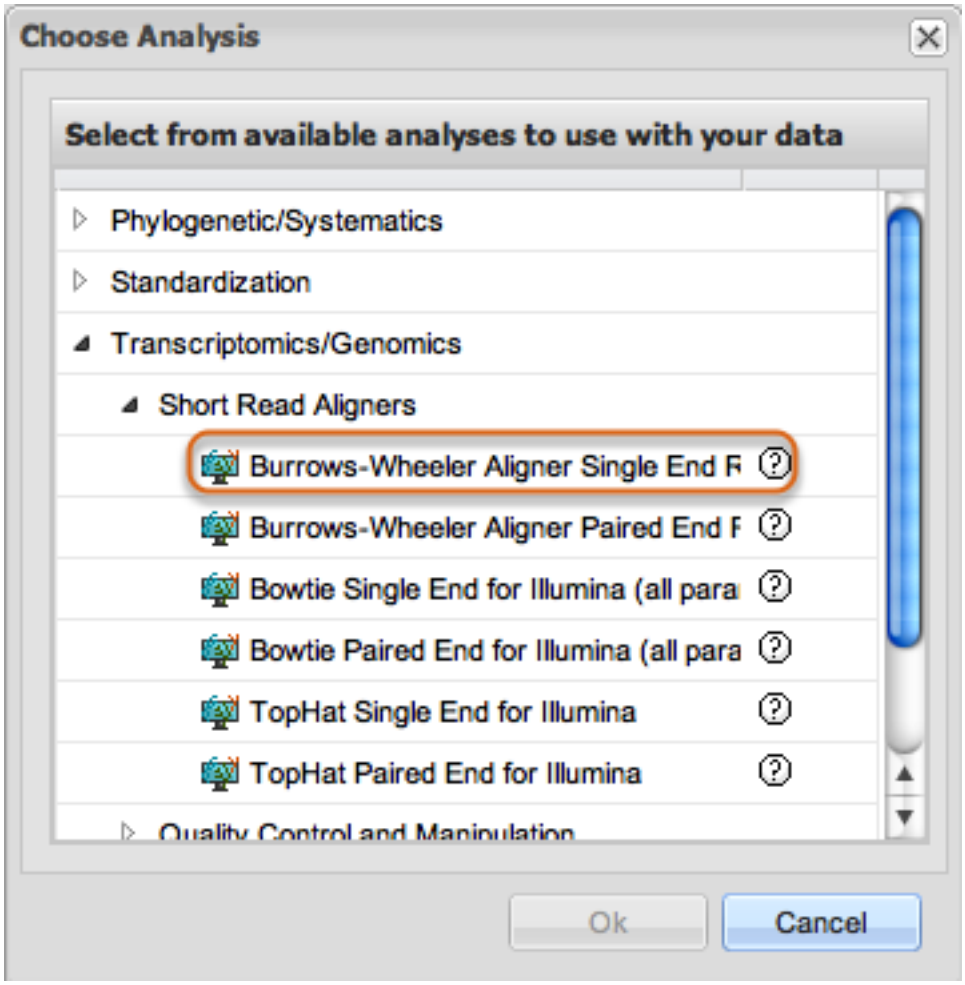
You may change the initial value for ML estimation or leave the default value in place.

Click Launch Job.

Enter a name and description for the job and click Ok. See Perform Analyses for information about monitoring the process and where to find your results.

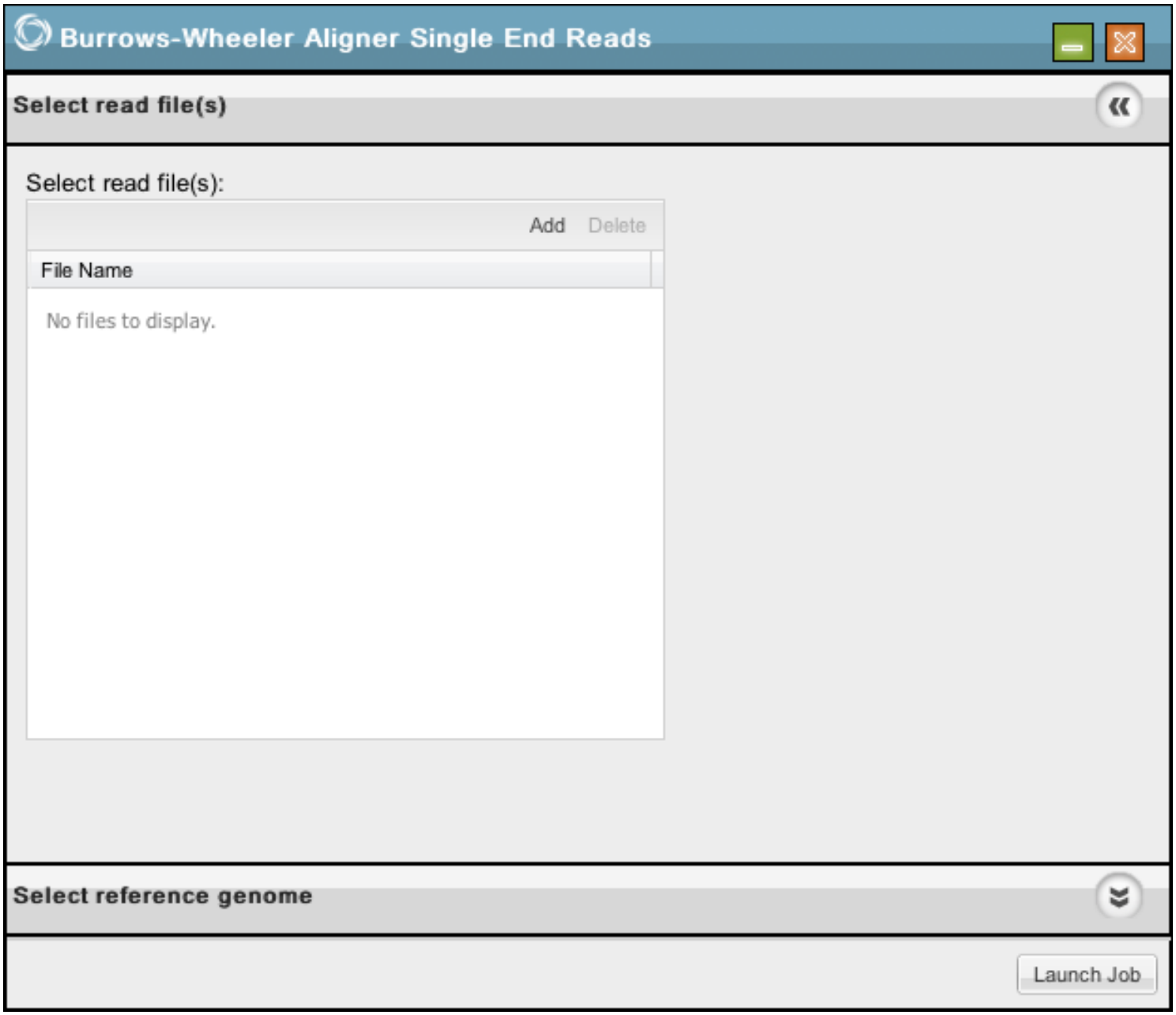
Burrows-Wheeler Aligner Single End Reads

This analysis uses the Burrows-Wheeler Aligner.



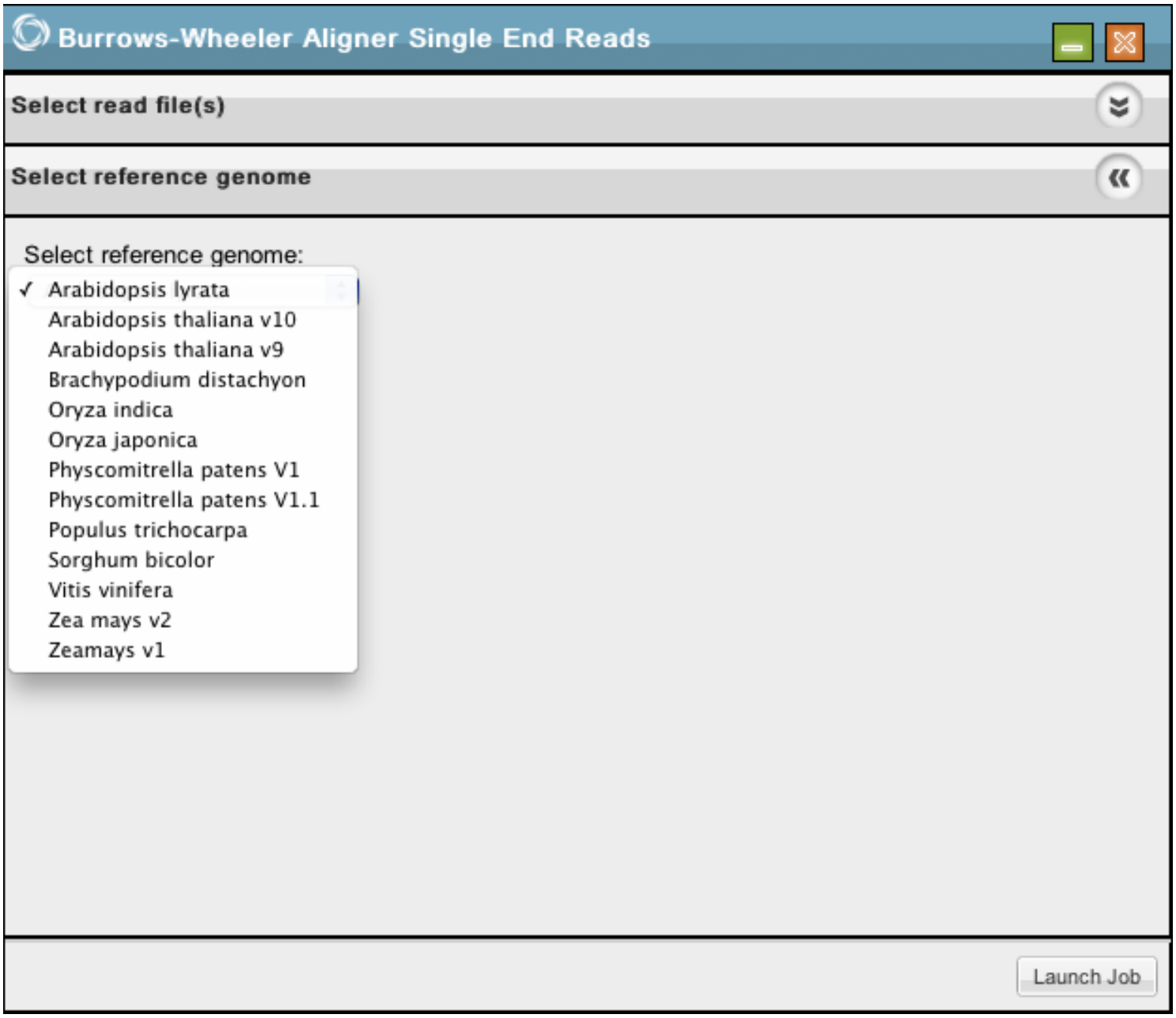
Select Burrows-Wheeler Aligner Single End Reads from within Perform Analyses as described in that section. Click Ok. An analysis is available for paired end reads.

Select reads



Click Add to select the previously uploaded and preprocessed DNA sequence read file that you want to align to a reference genome.

Select reference genome



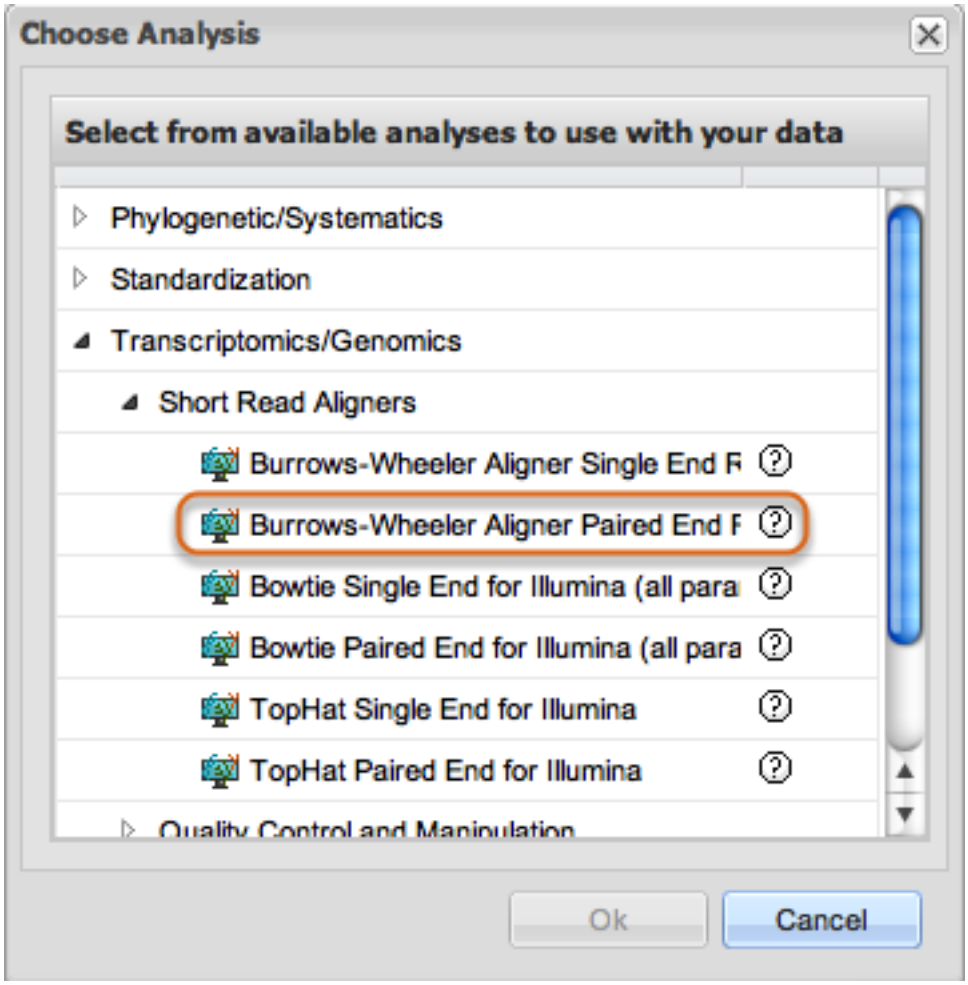
Click the arrow to open a drop-down box listing available reference genomes. Click one to select it.

Click Launch Job.

Enter a name and description for the job and click Ok. See Perform Analyses for information about monitoring the process and where to find your results.

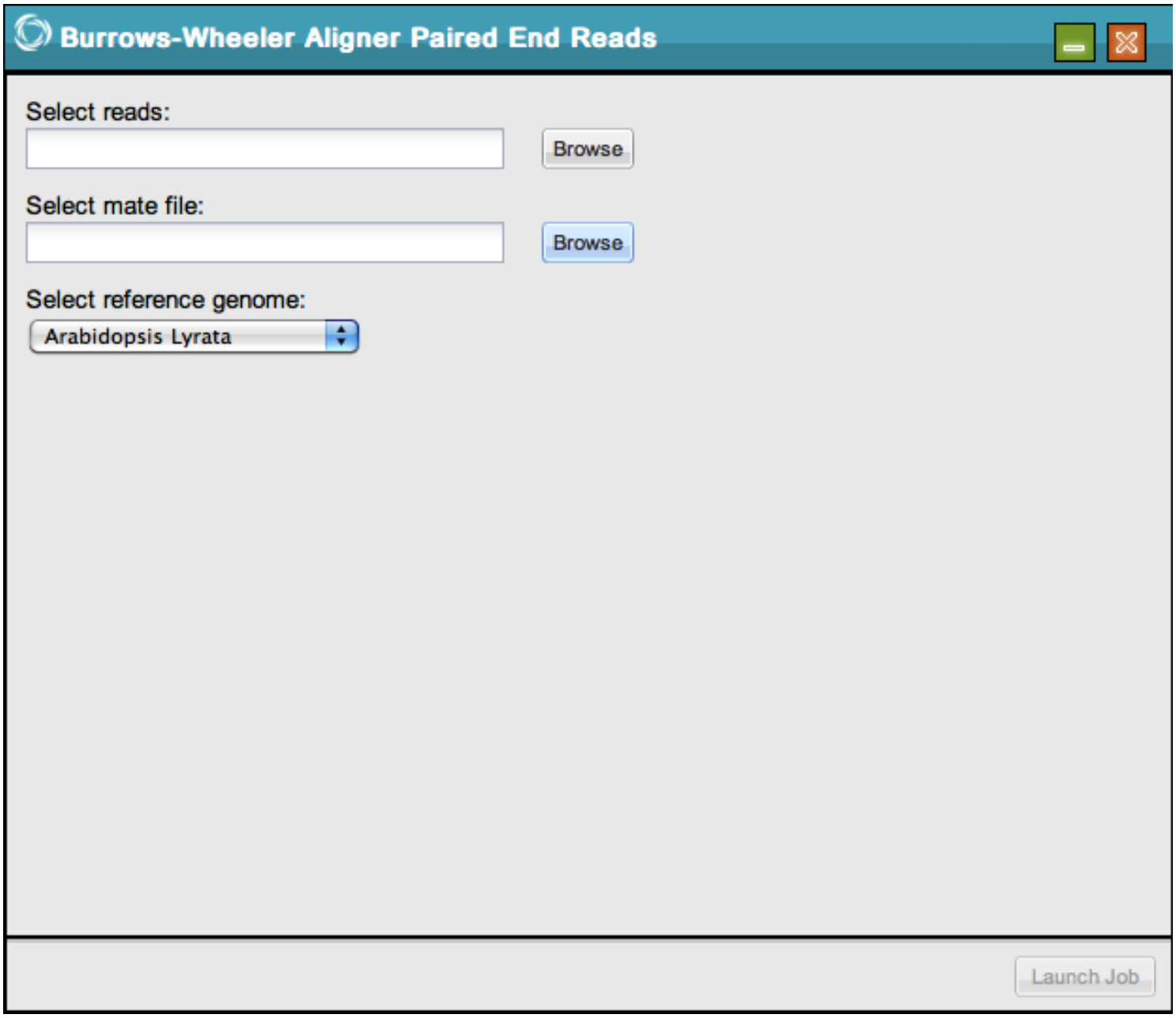
Burrows-Wheeler Aligner Paired End Reads

This analysis uses the Burrows-Wheeler Aligner.



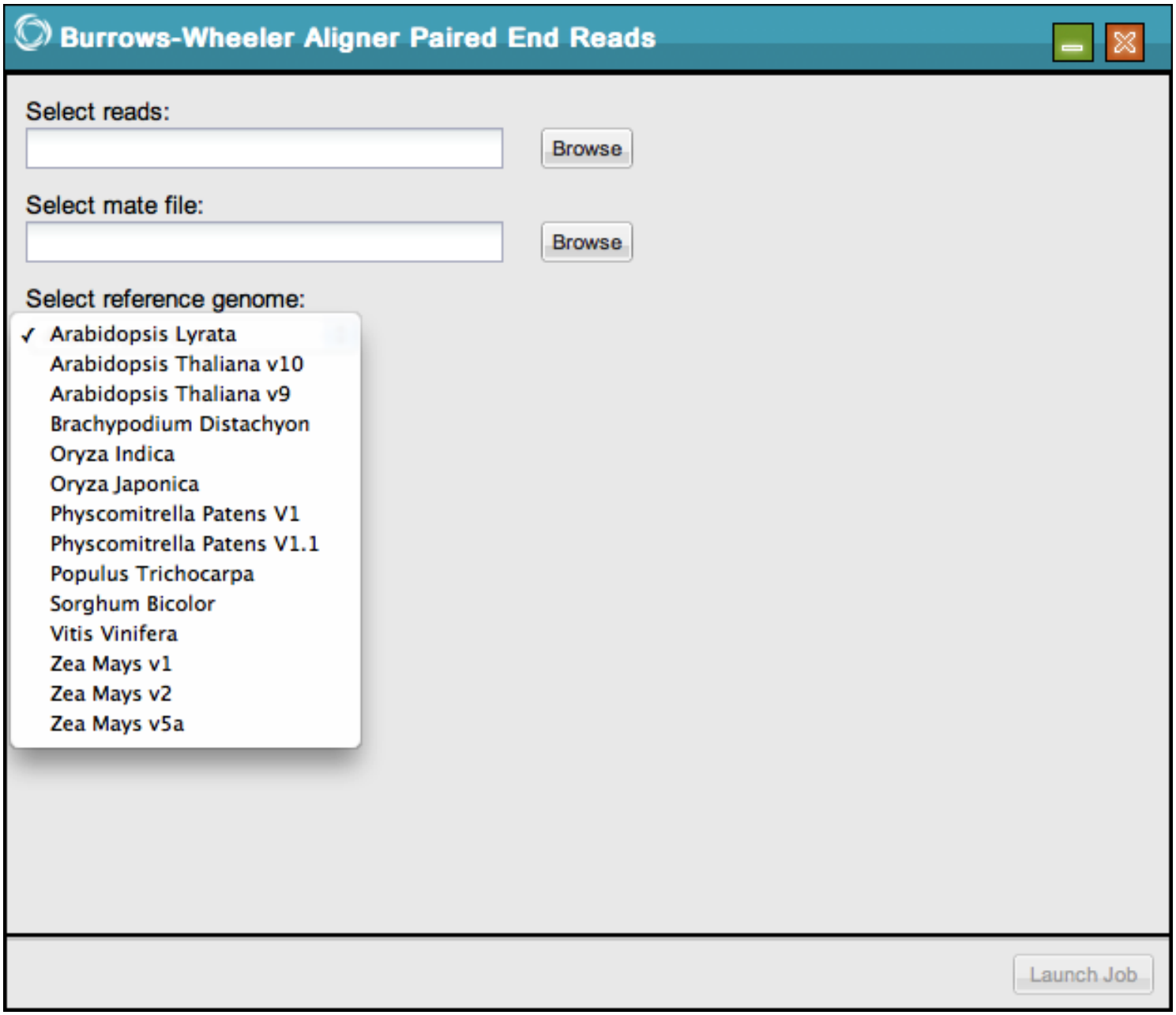
Select Burrows Wheeler Aligner Paired End Reads from within Perform Analyses as described in that section. Click Ok. An analysis is available for single end reads.

Select reads



Click Browse next to Select reads and Select mate file to select the previously uploaded and preprocessed DNA sequence read file and mate file that you want to align to a reference genome.

Select reference genome



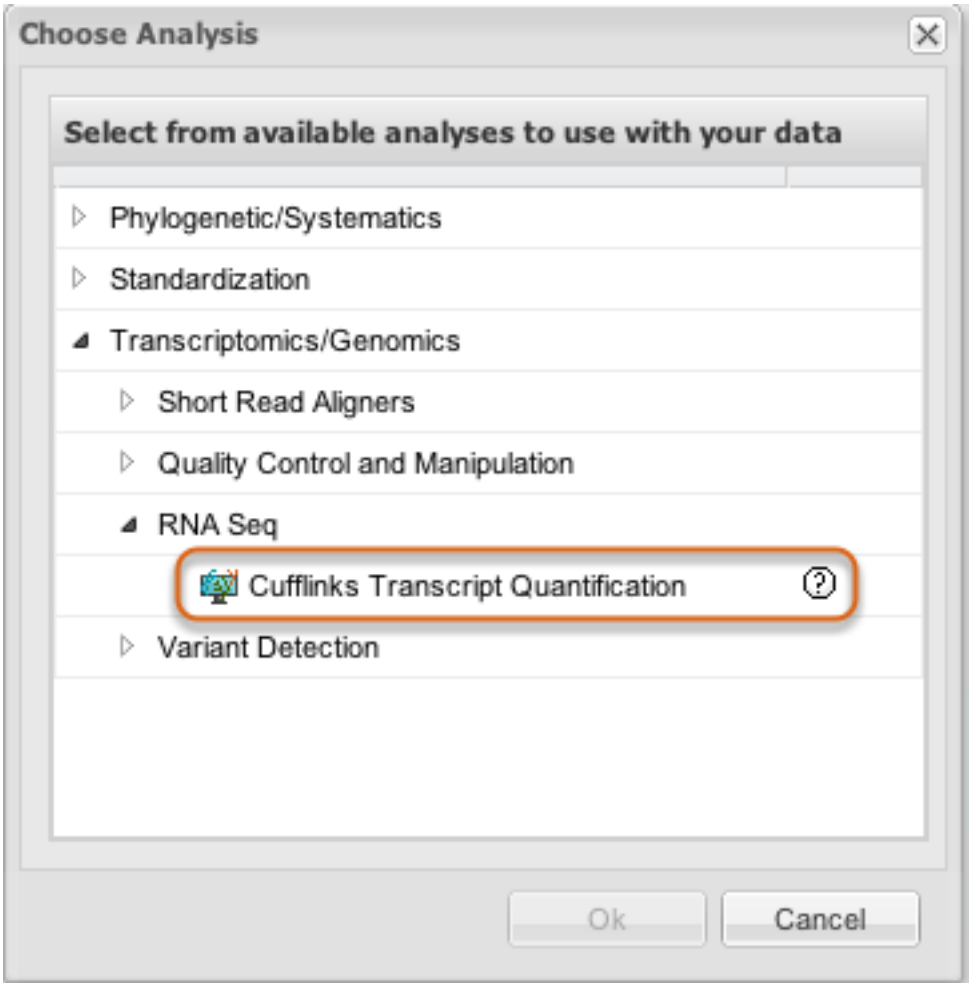
Click the arrow to open a drop-down box listing available reference genomes. Click one to select it.

Click Launch Job.

Enter a name and description for the job and click Ok. See Perform Analyses for information about monitoring the process and where to find your results.

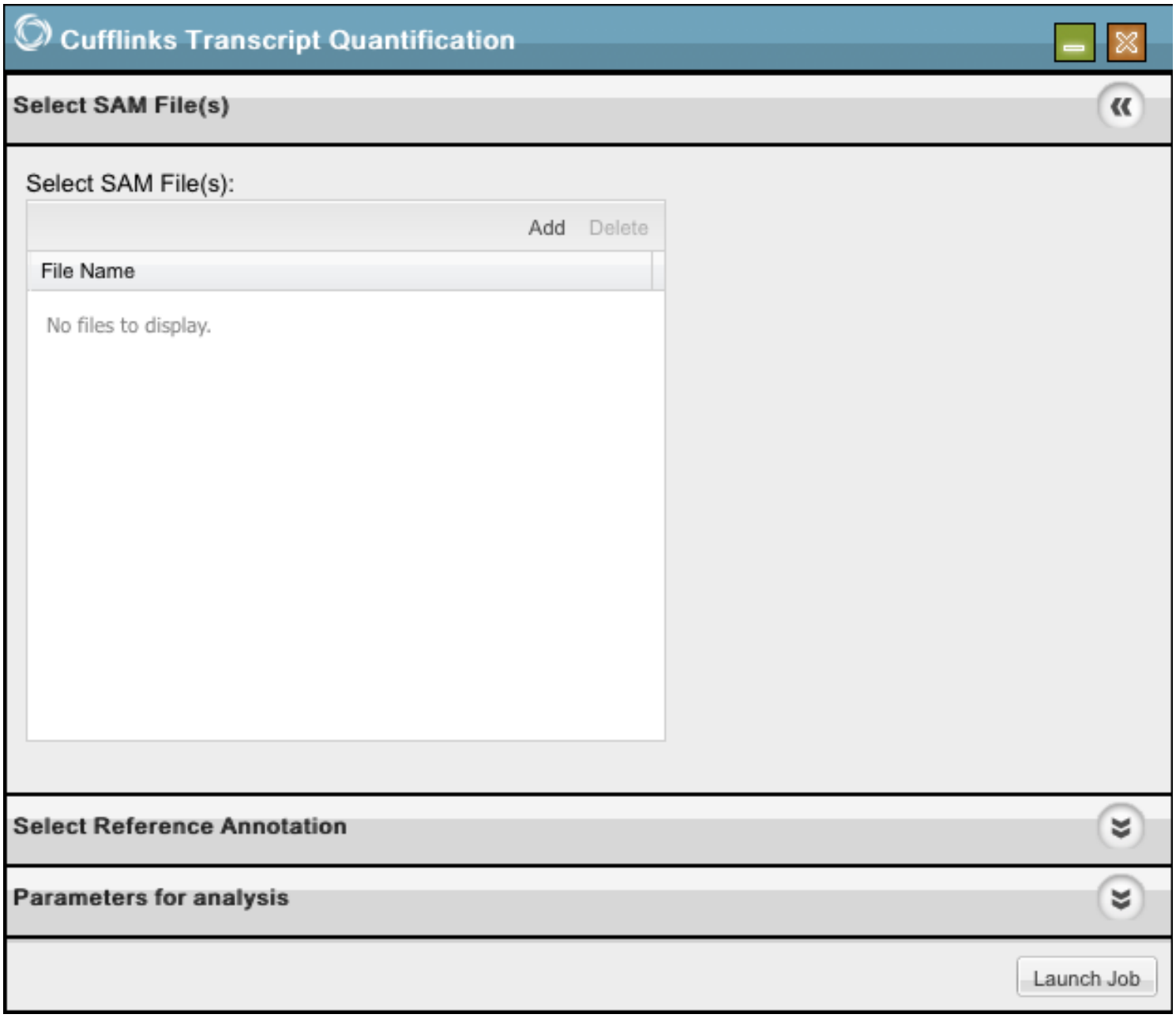
Cufflinks Transcript Quantification

This analysis uses Cufflinks.



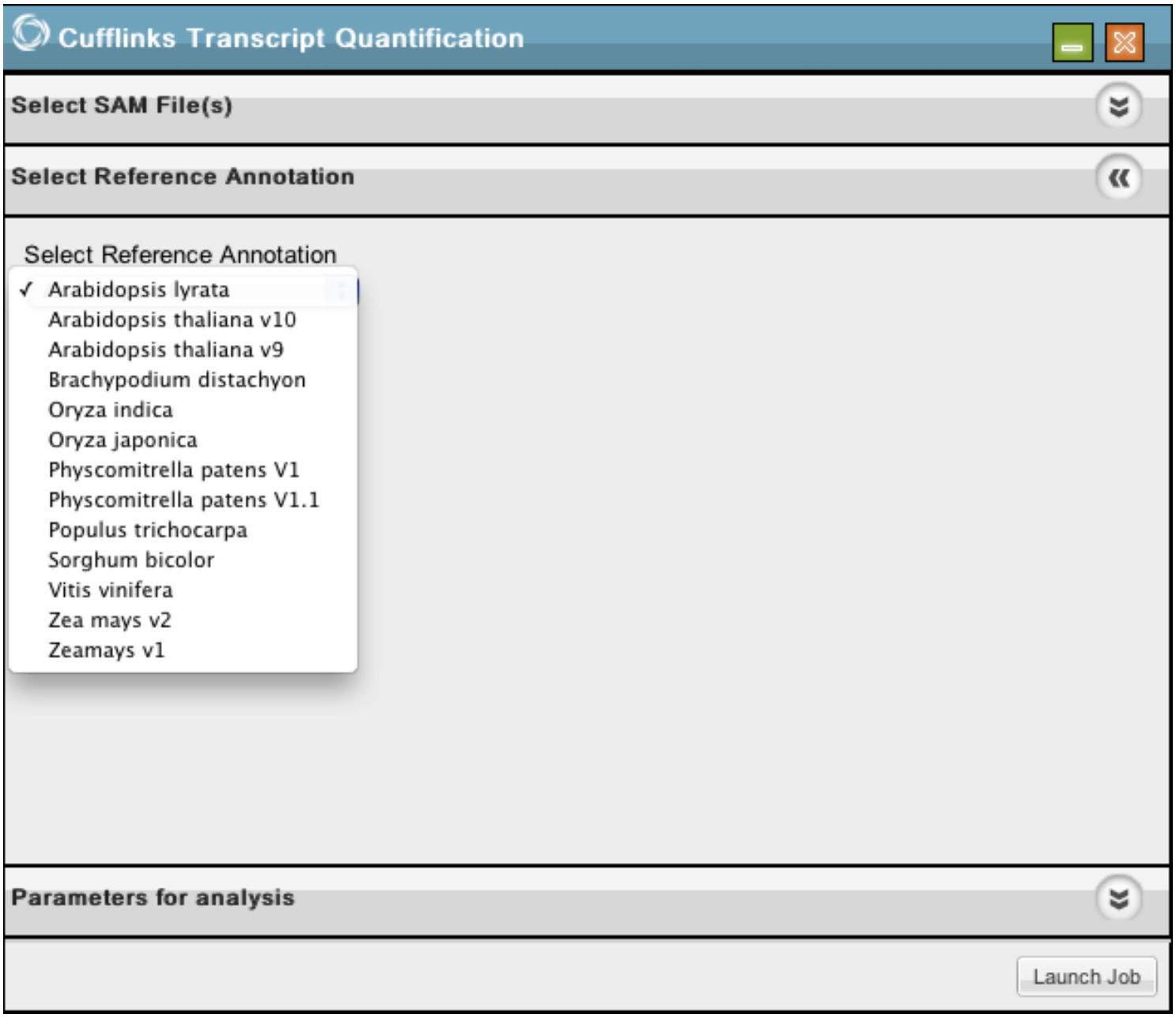
Select Cufflinks Transcript Quantification from within Perform Analyses as described in that section. Click Ok.

Select SAM File(s)



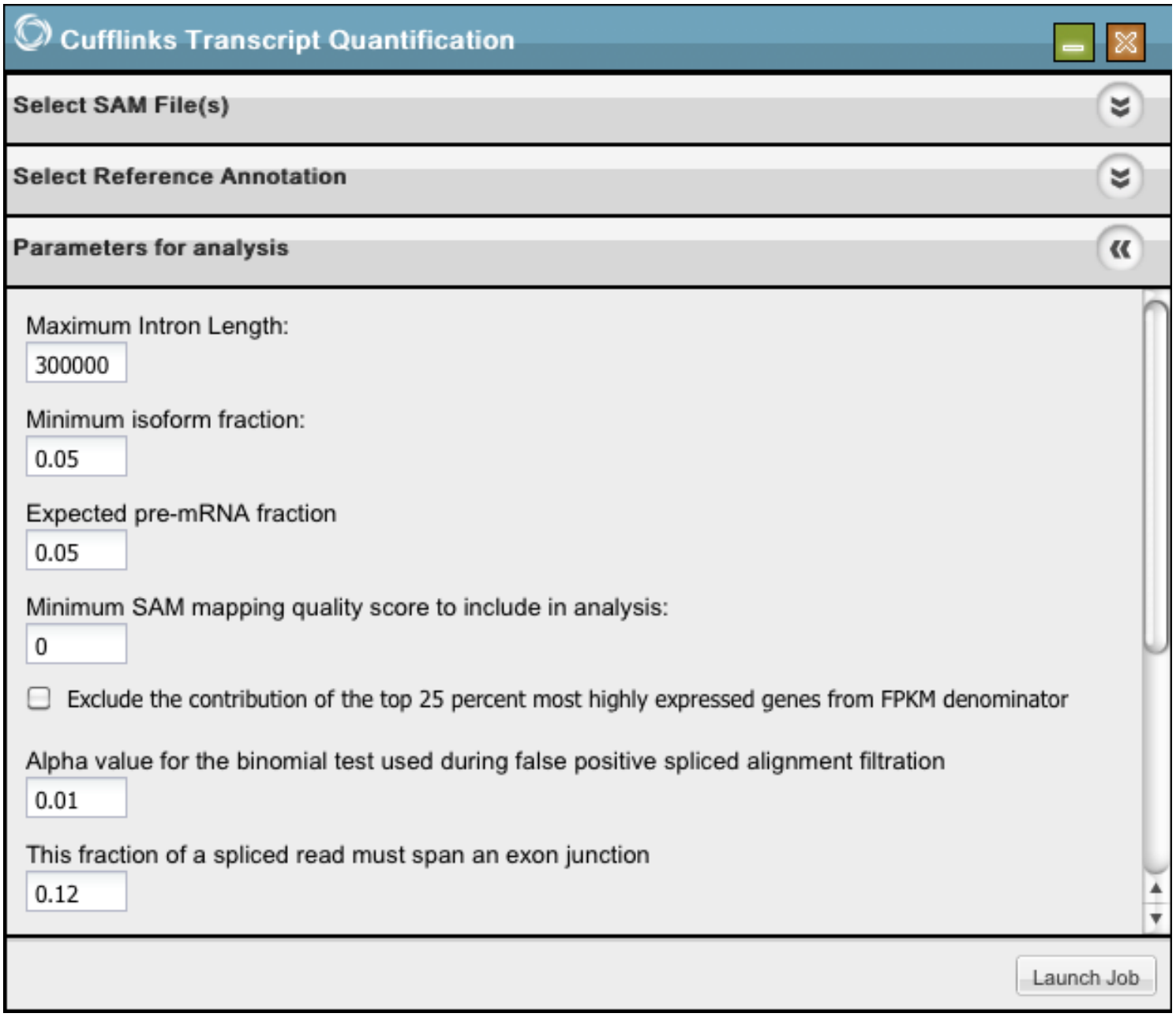
Click Add to choose your previously uploaded SAM file(s).

Select Reference Annotation



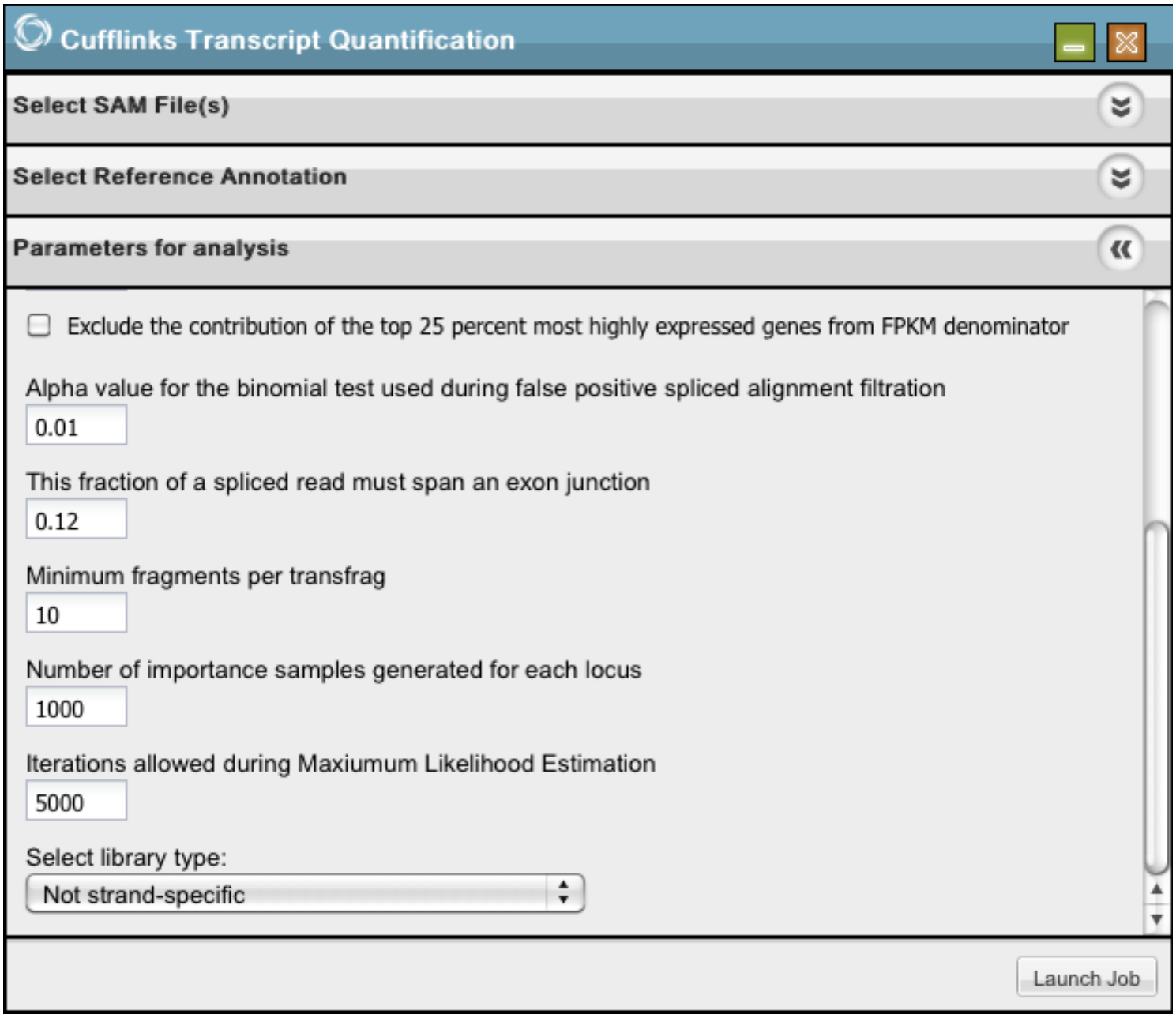
Select the reference genome.

Parameters (part one)


Cufflinks Transcript Quantification
Select SAM File(s)
Select Reference Annotation
Parameters for analysis
Maximum Intron Length:
300000
Minimum isoform fraction:
0.05
Expected pre-mRNA fraction
0.05
Minimum SAM mapping quality score to include in analysis:
0
Exclude the contribution of the top 25 percent most highly expressed genes from FPKM denominator
Alpha value for the binomial test used during false positive spliced alignment filtration
0.01
This fraction of a spliced read must span an exon junction
0.12
Launch Job

Select your desired parameters (continued in following image).

Parameters (part two)

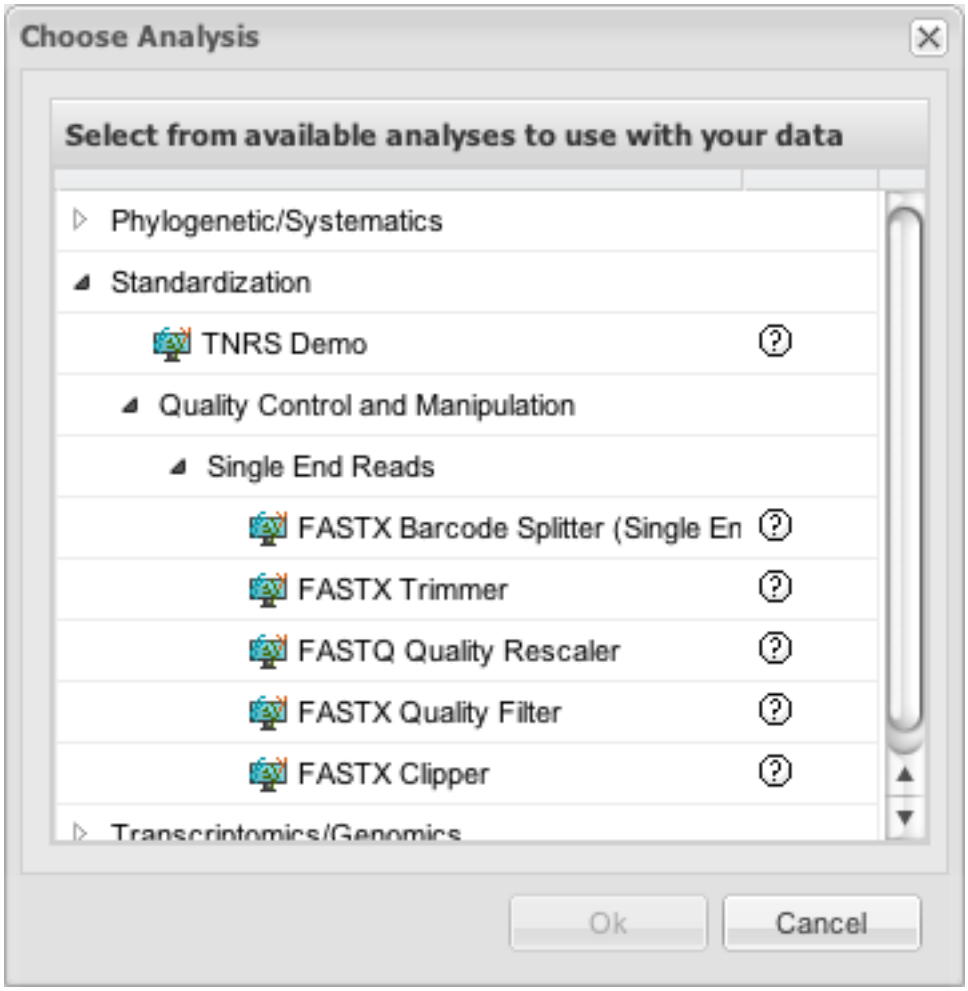

Cufflinks Transcript Quantification
Select SAM File(s)
Select Reference Annotation
Parameters for analysis
Exclude the contribution of the top 25 percent most highly expressed genes from FPKM denominator
Alpha value for the binomial test used during false positive spliced alignment filtration
0.01
This fraction of a spliced read must span an exon junction
0.12
Minimum fragments per transfrag
10
Number of importance samples generated for each locus
1000
Iterations allowed during Maxiumum Likelihood Estimation
5000
Select library type:
Not strand-specific
Launch Job

Click Launch Job.

Enter a name and description for the job and click Ok.

See Perform Analyses for information about monitoring the process and where to find your results.

FASTX Analyses Overview



The FASTX-Toolkit is a collection of command line tools for preprocessing of DNA and RNAseq Short-Reads. Several of these are available as analyses in the Discovery Environment. They are found in Perform Analyses under Choose Analysis.

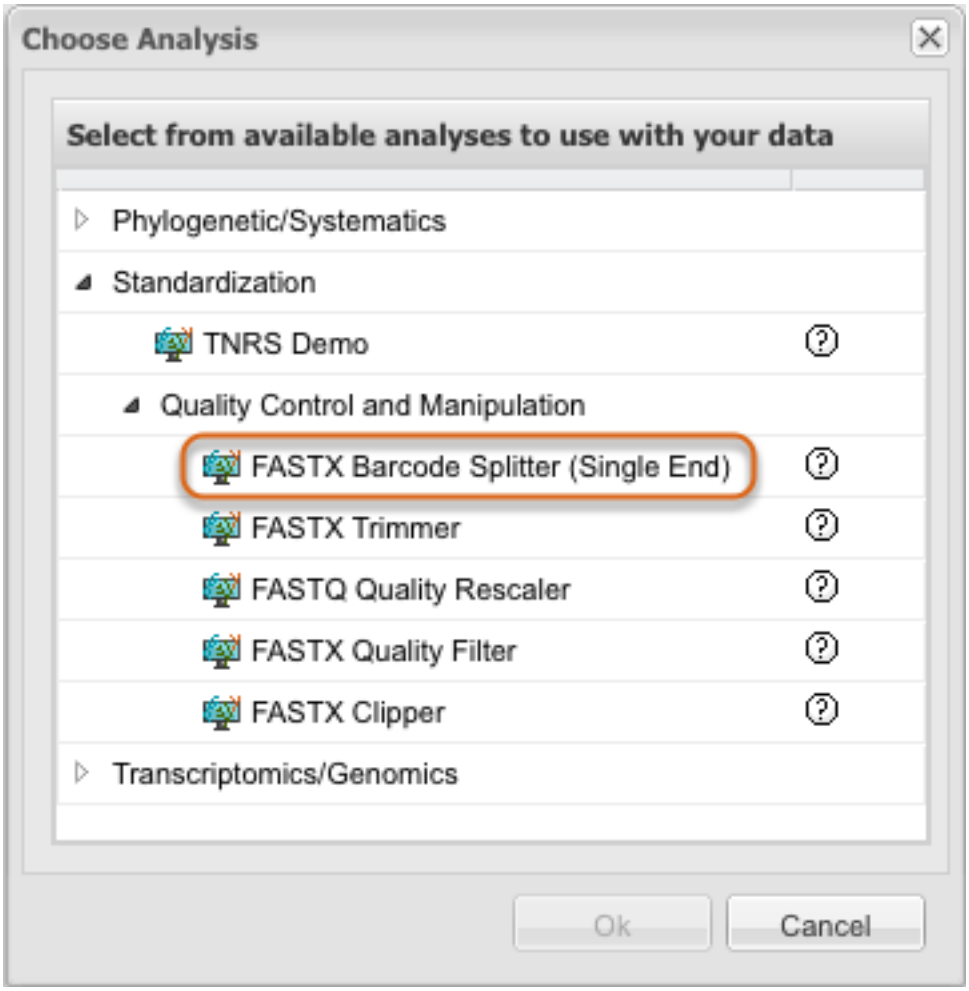
Each of these is described in a separate section.

FASTX Barcode Splitter (Single End)

An overview of FASTX Analyses is available.

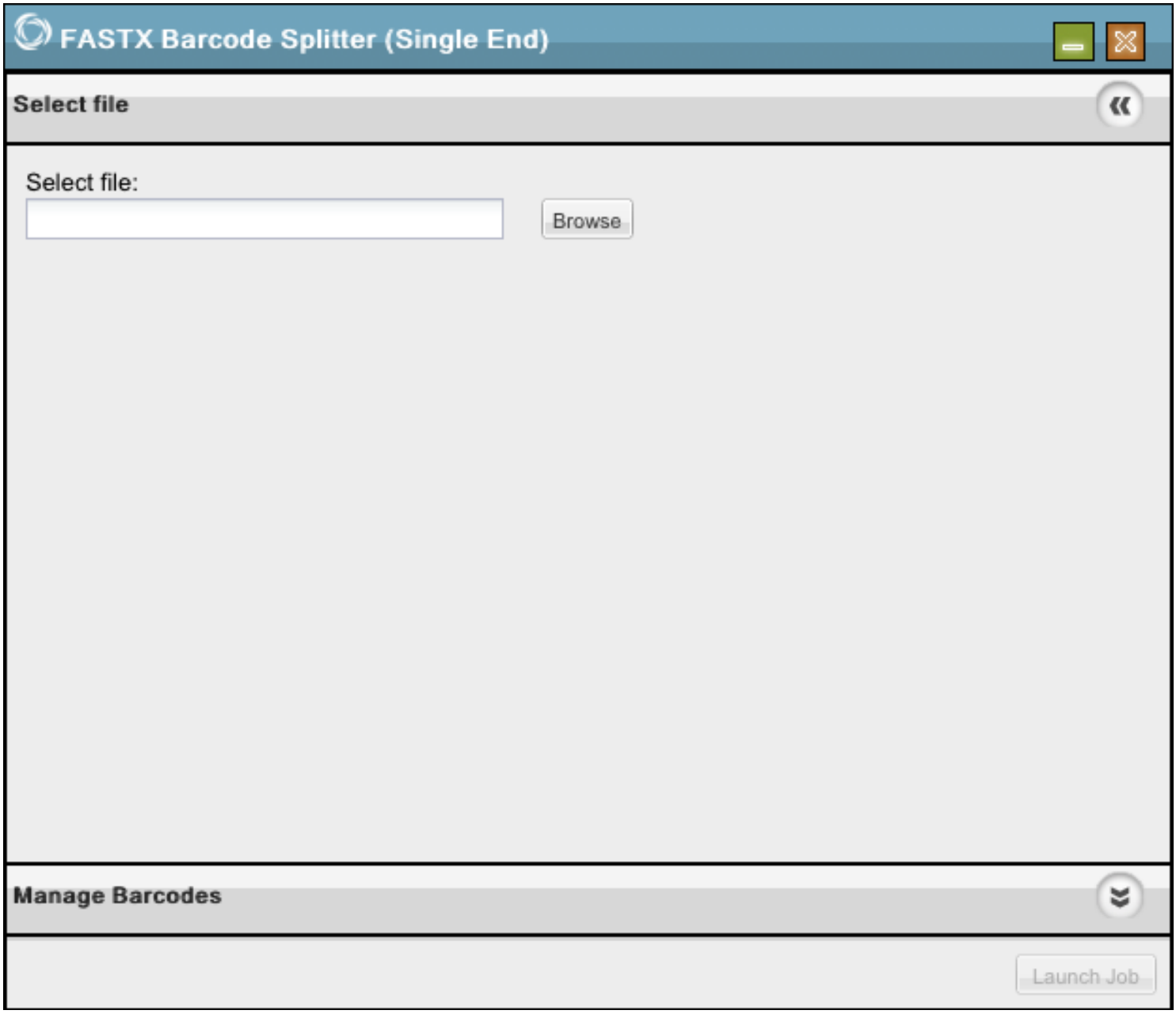
The FASTX Barcode Splitter splits a FASTQ file into several files using barcodes as the split criteria.

Barcode files are simple text files. Each line should contain an identifier (descriptive name for the barcode), and the barcode itself (A/C/G/T), separated by a TAB character or a space. An example is given in an image on the FASTX documentation website.



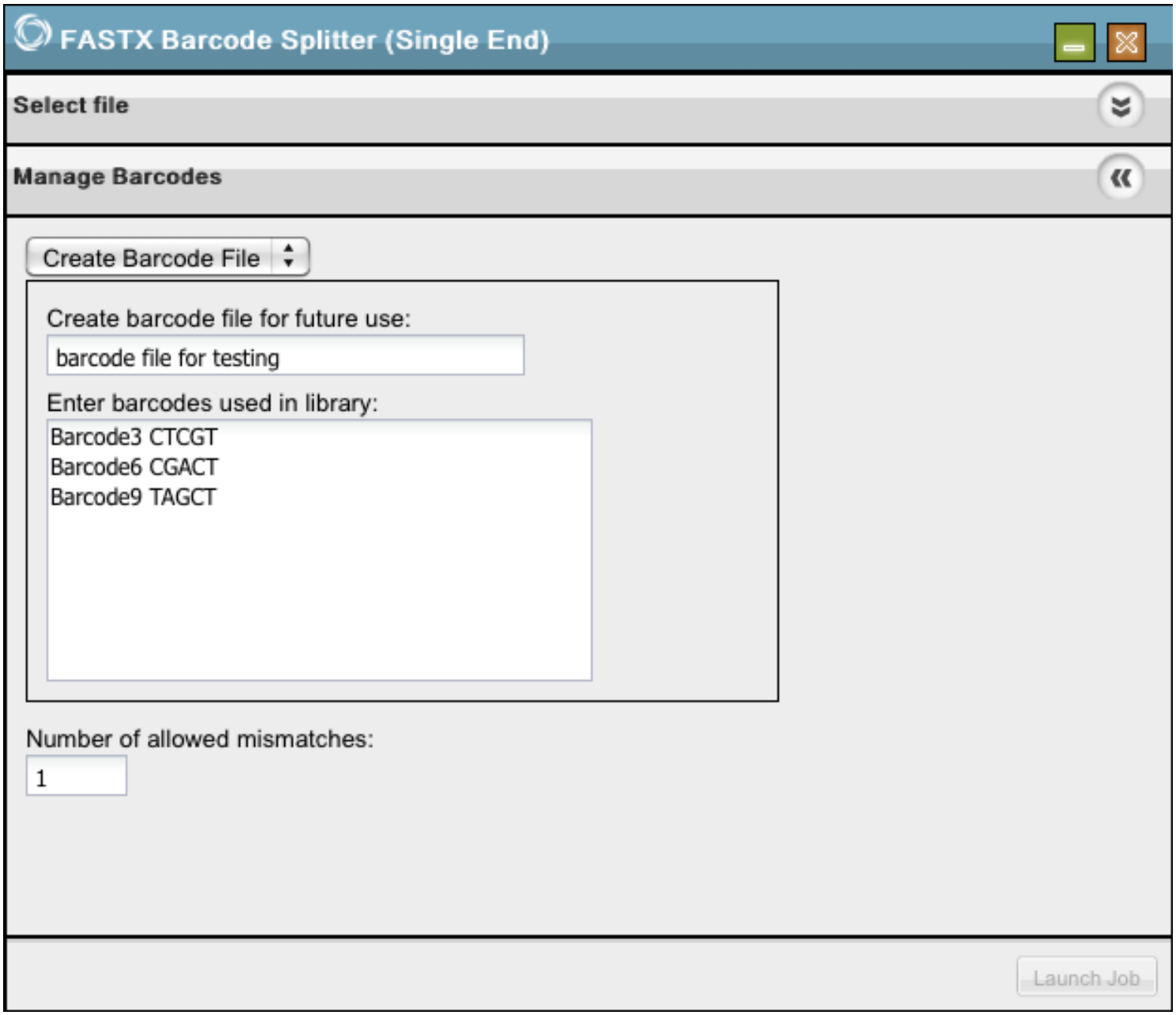
Select FASTX Barcode Splitter from within Perform Analyses as described in that section. Click Ok.

Select file



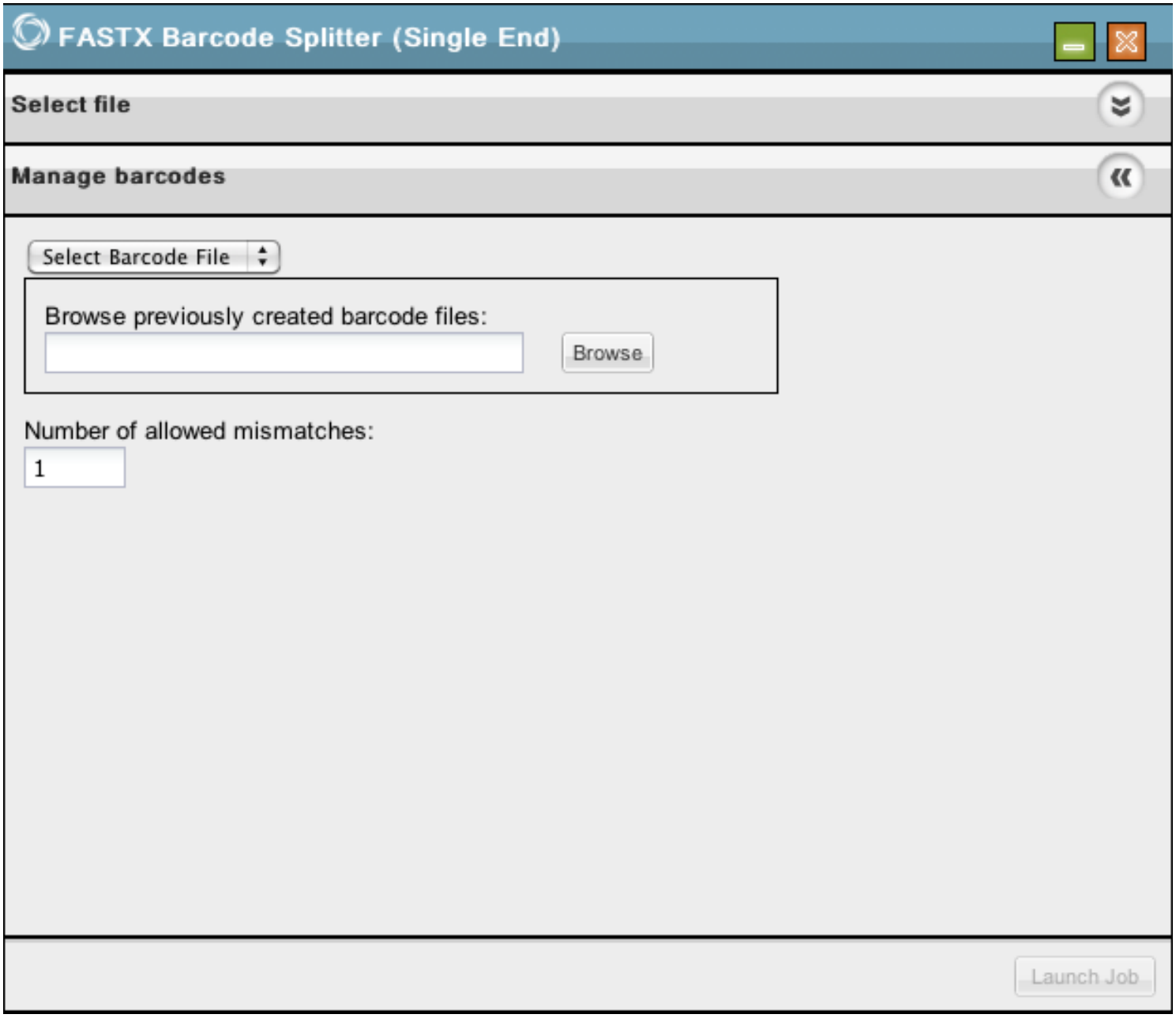
Click Browse to select your previously uploaded file. Click Manage Barcodes.

Manage Barcodes, create



Choose Create Barcode File from the drop-down menu if you are going to create one now. Create a name for the file to help you locate it later. Enter your barcodes, each on a new line, separate titles and codes with a space. Click Launch Job.

Manage Barcodes, select

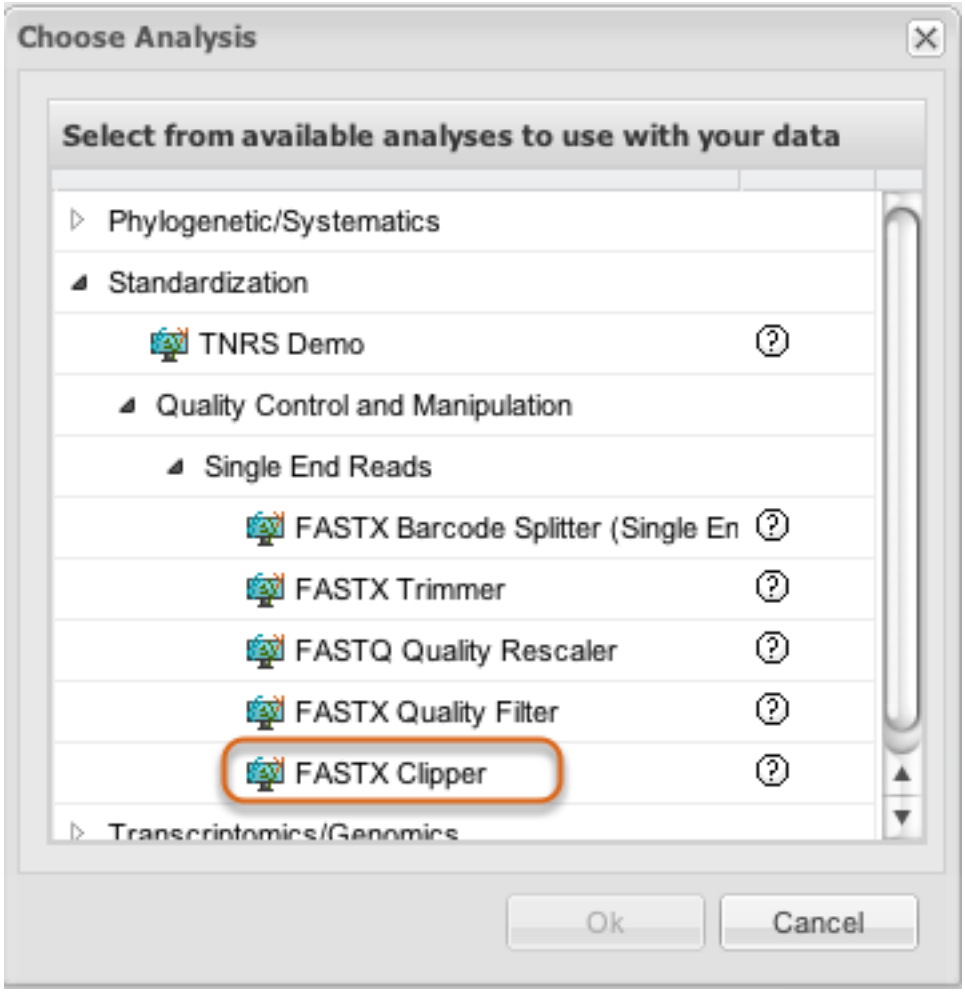


Choose Select Barcode File if you have previously uploaded one to the Discovery Environment.
Click Launch Job.

Enter a name and description for the job and click Ok. See Perform Analyses for information about monitoring the process and where to find your results.

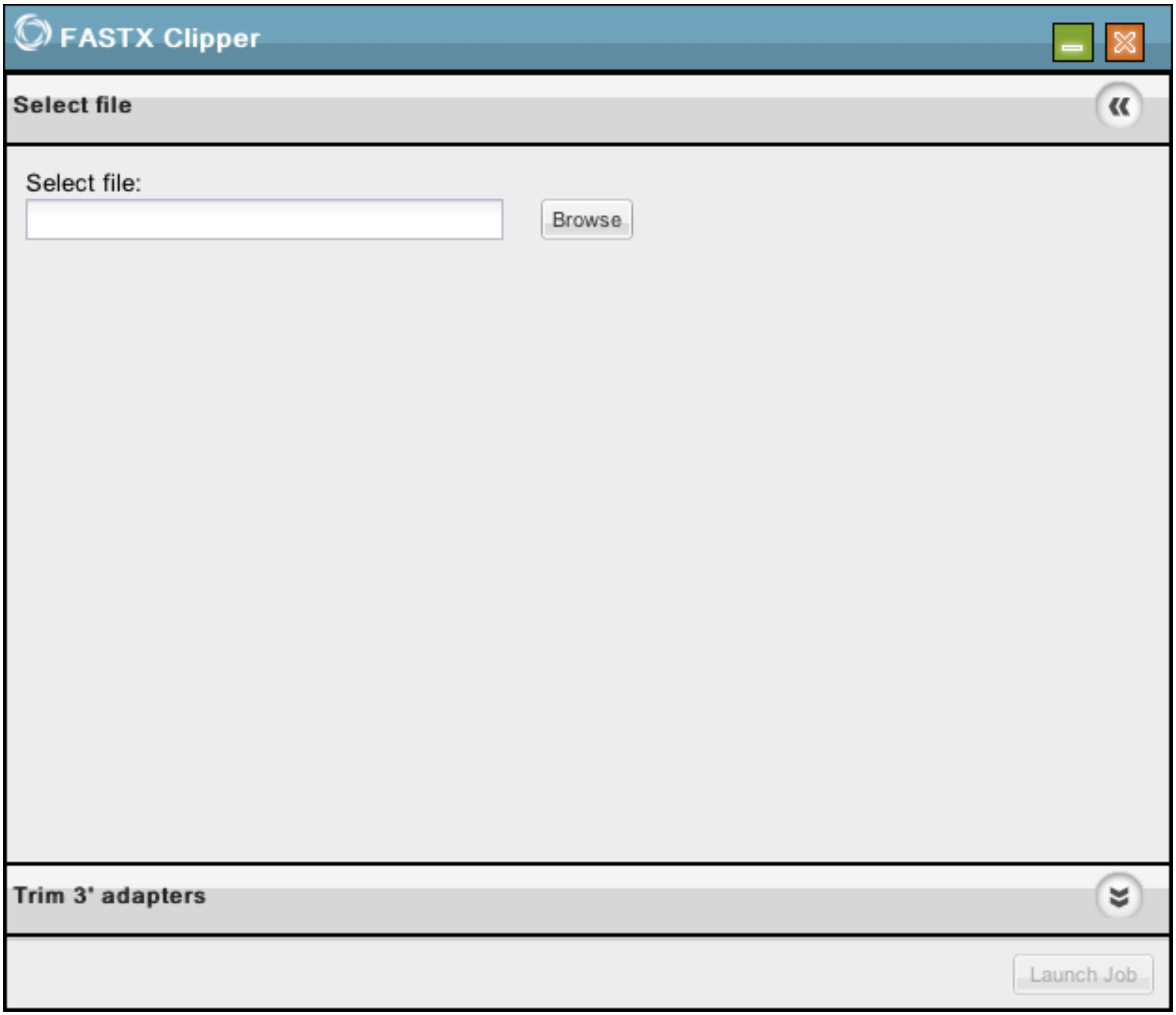
FASTX Clipper

An overview of FASTX Analyses is available.

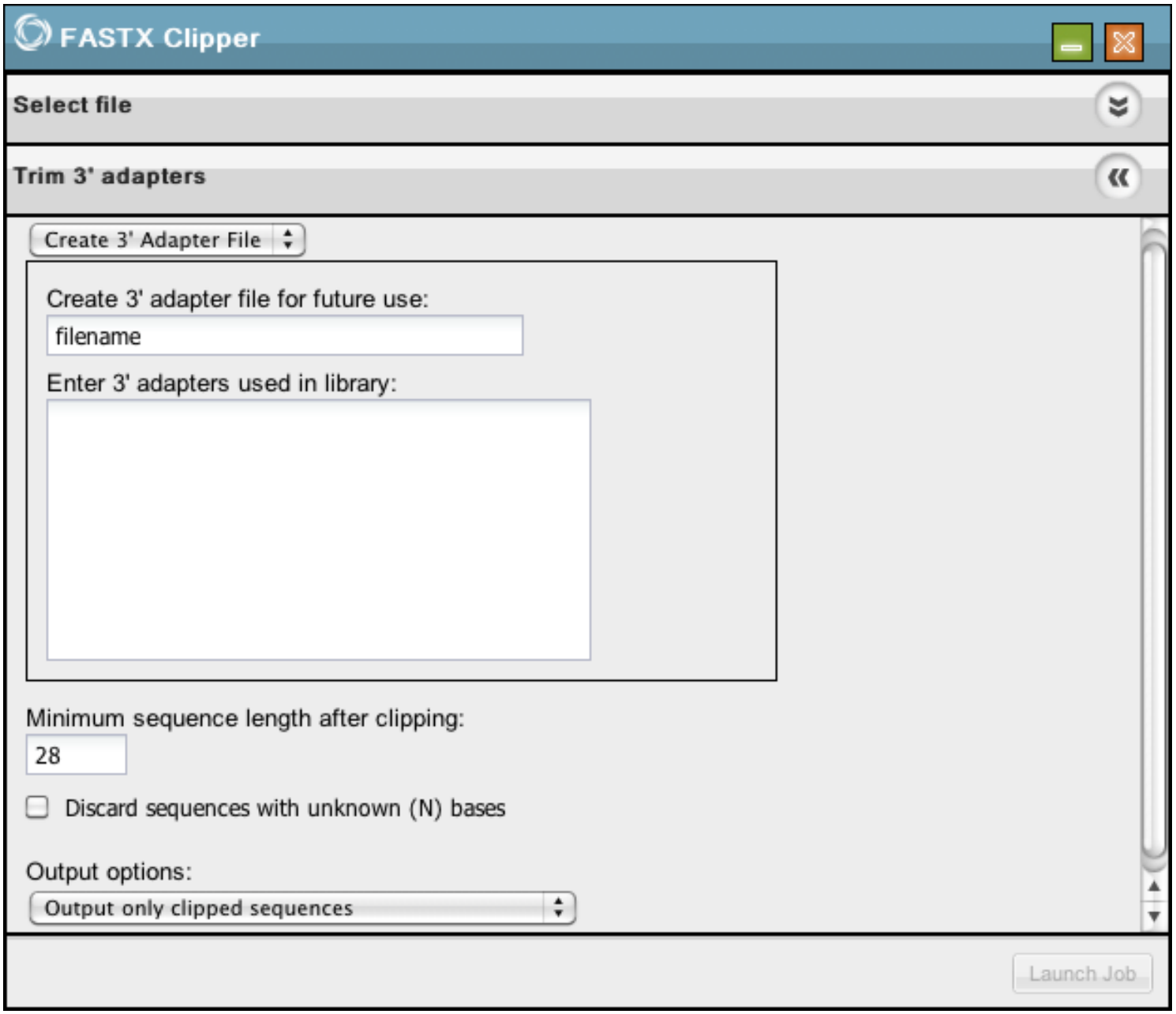


Select FASTX Clipper from within Perform Analyses as described in that section. Click Ok.

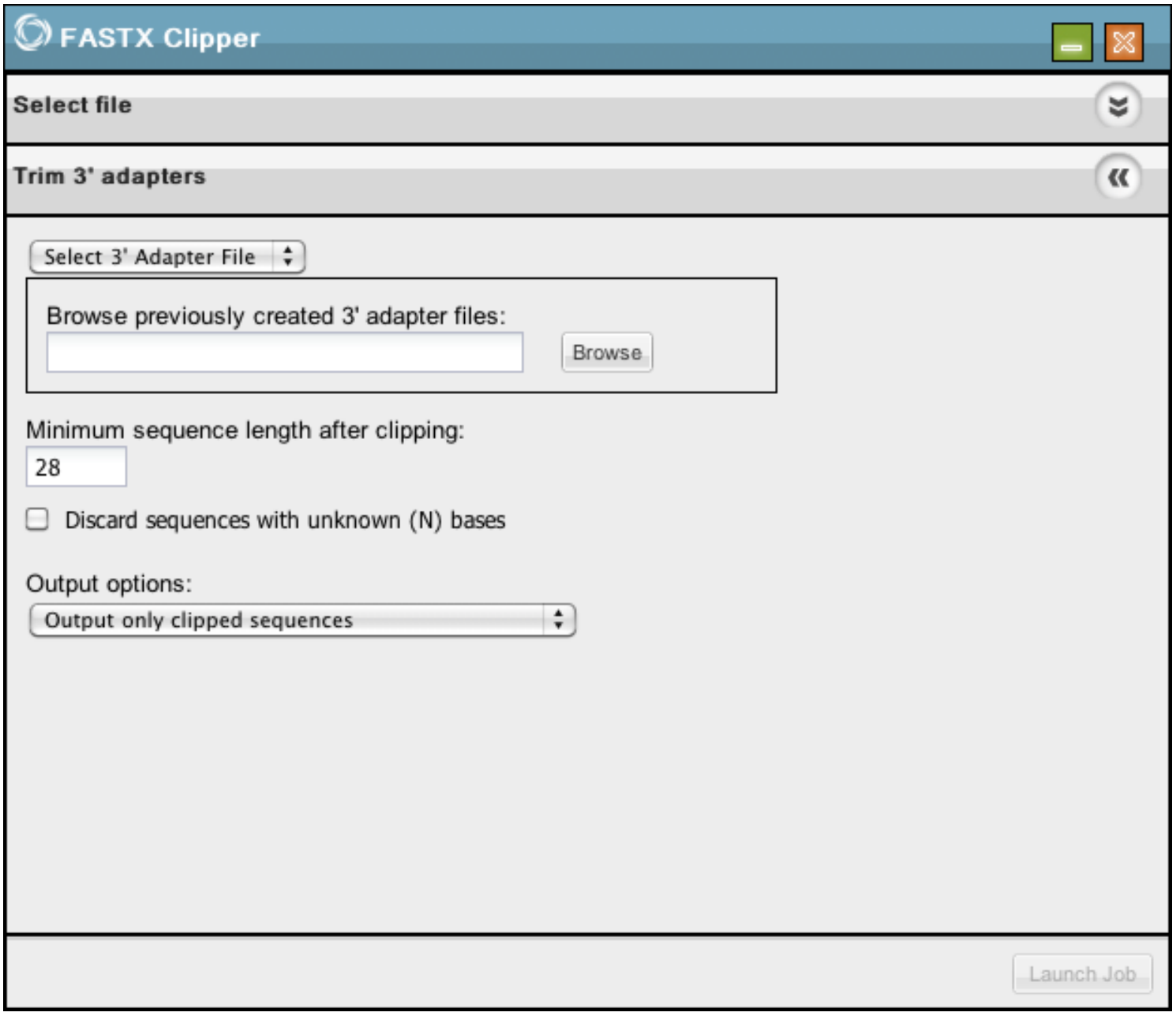
Single end read data input



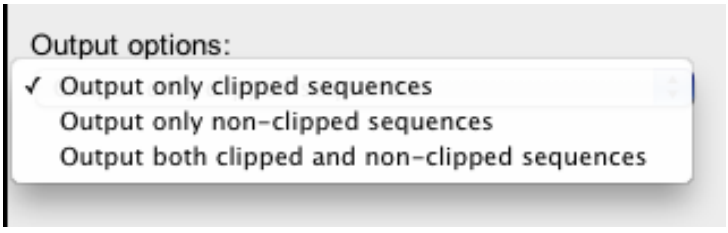
Click Browse to select your previously uploaded file. Click Trim 3' Adapters.



Choose Create 3' Adapter File from the drop-down menu if you are going to create one now.



Choose Select 3' Adapter File from the drop-down menu if you are going to use a previously uploaded file.

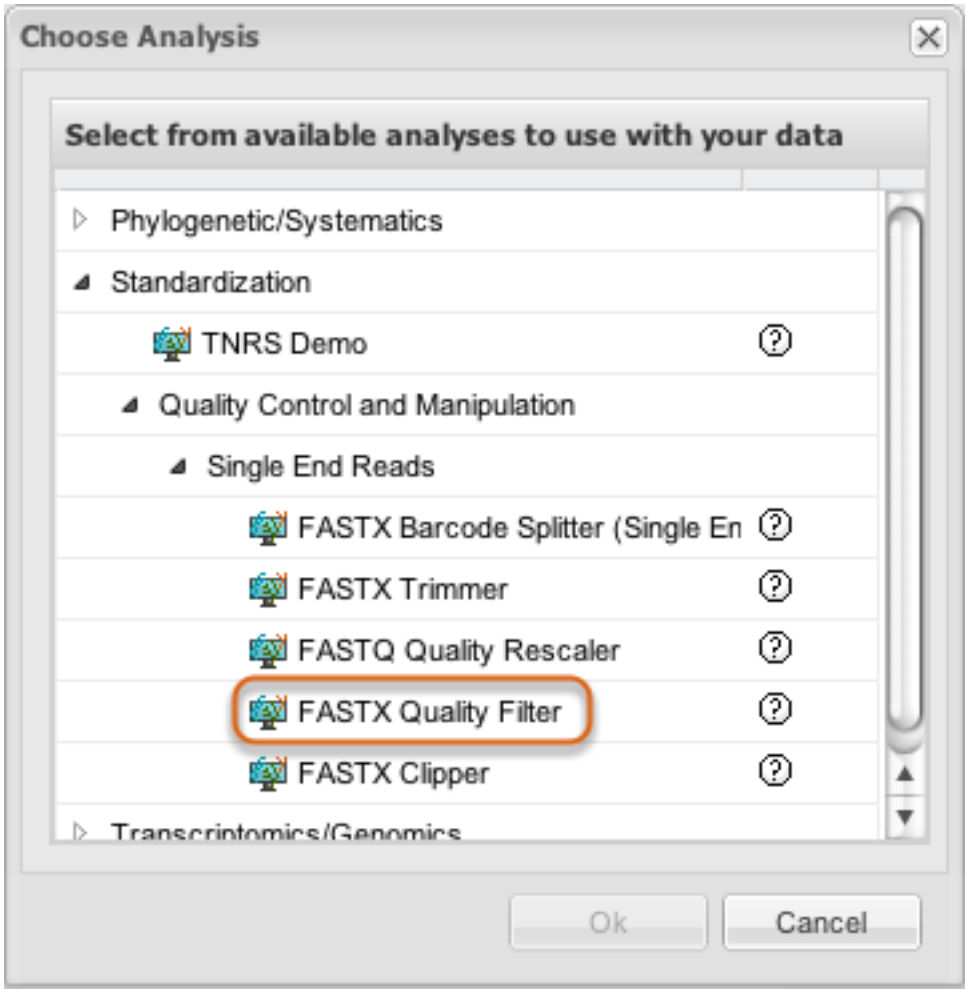

Output options:
✓ Output only clipped sequences
Output only non-clipped sequences
Output both clipped and non-clipped sequences

Keep or modify the default settings. Choose your desired output option from the Output options drop-down menu. Click Launch Job.

Enter a name and description for the job and click Ok. See Perform Analyses for information about monitoring the process and where to find your results.

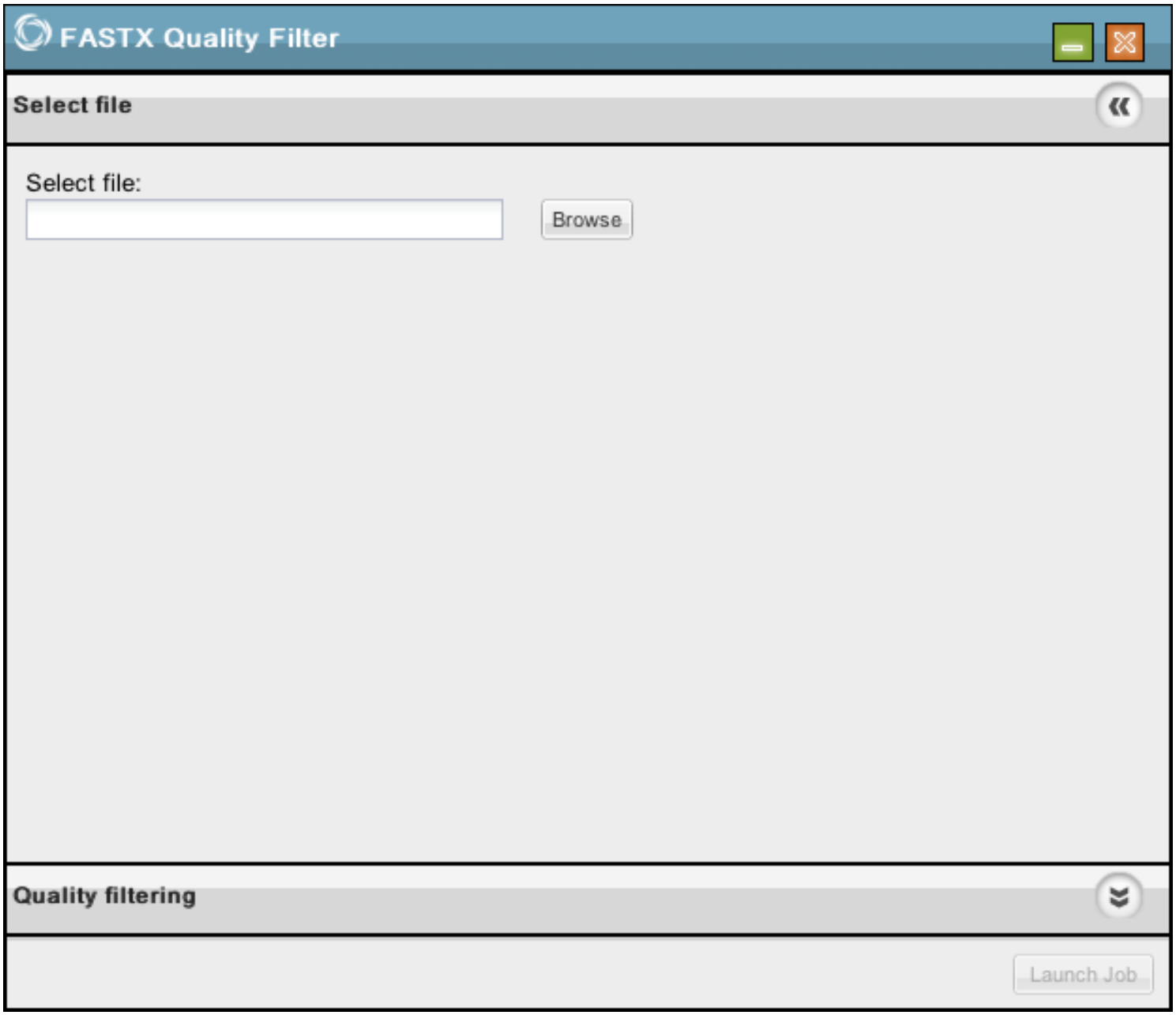
FASTX Quality Filter

An overview of FASTX Analyses is available.



Select FASTX Quality Filter from within Perform Analyses as described in that section. Click Ok.

Select file



Click Browse to select your previously uploaded file. Click Quality filtering.

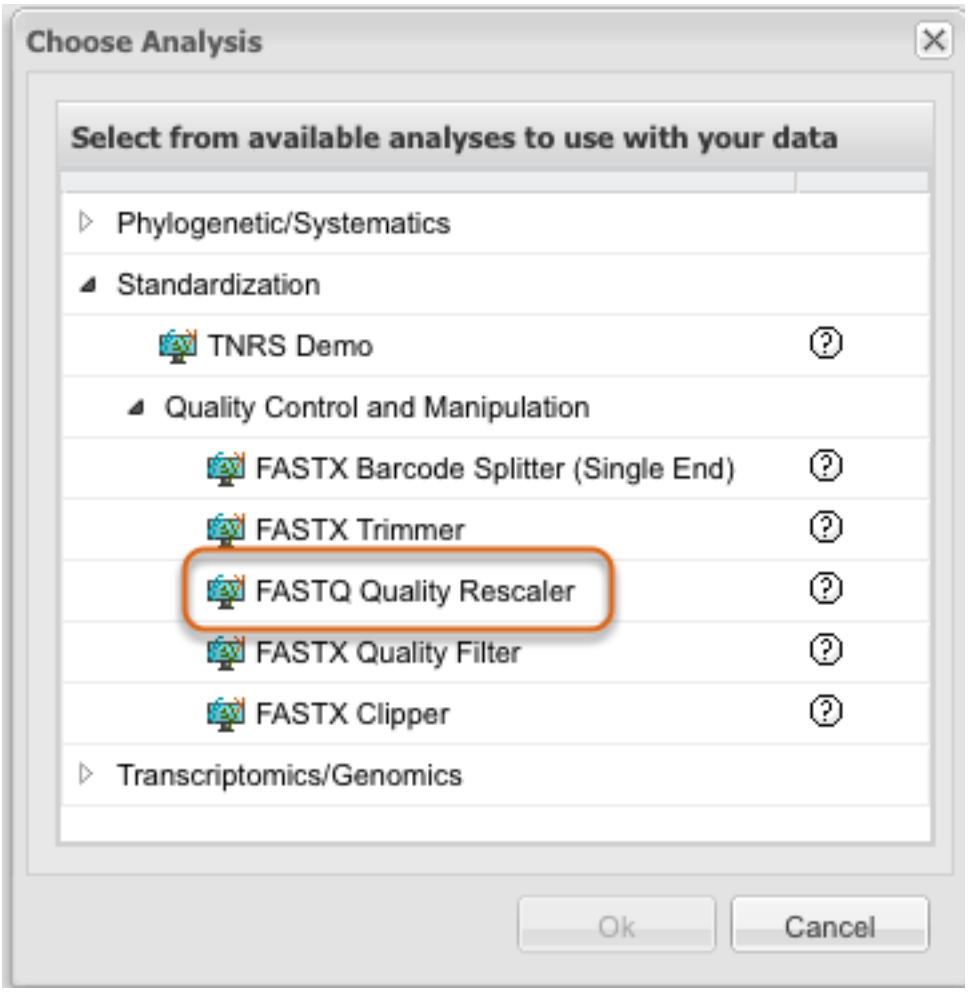
Keep or modify the default settings. Click Launch Job.

Enter a name and description for the job and click Ok. See Perform Analyses for information about monitoring the process and where to find your results.

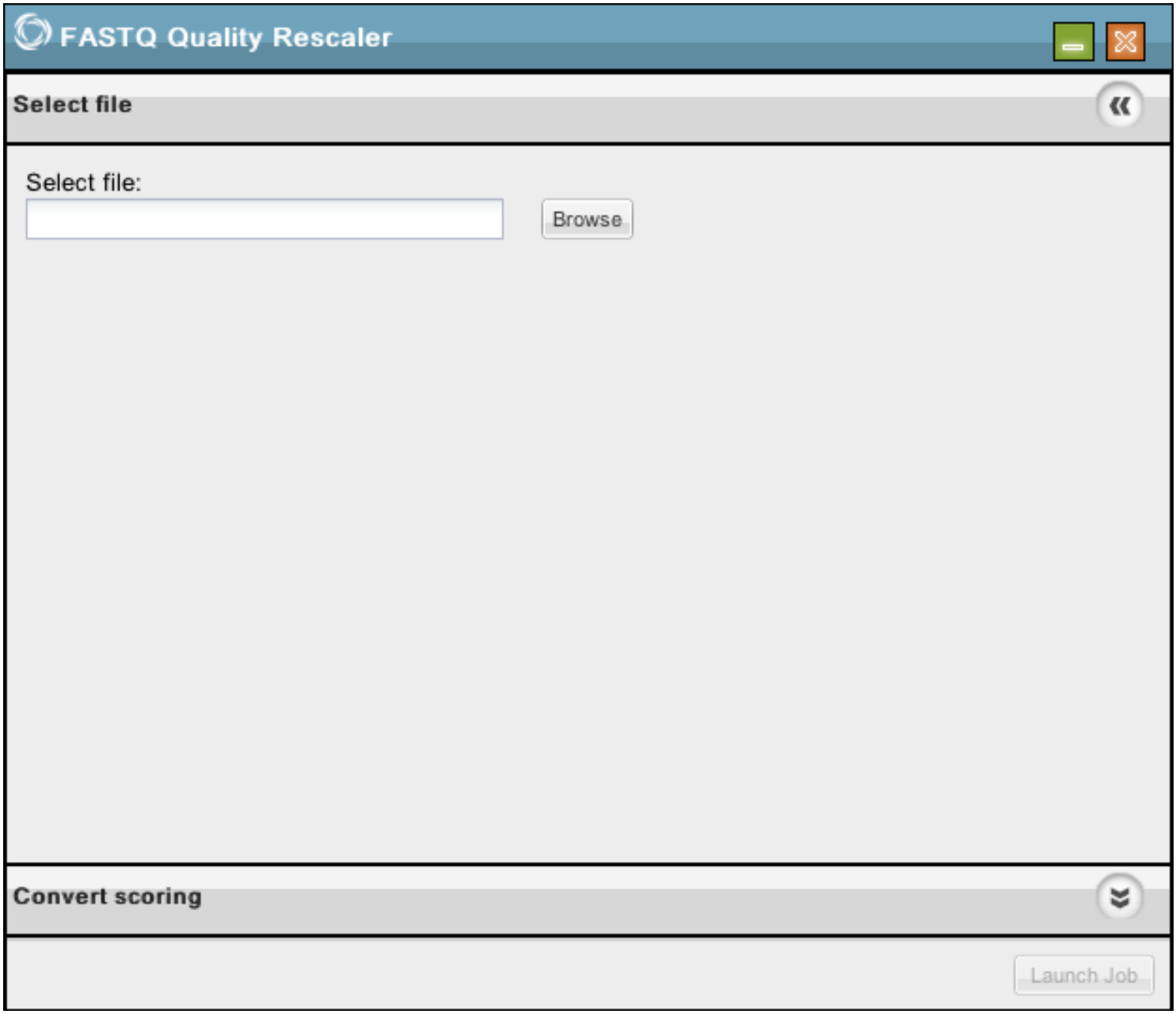
FASTQ Quality Rescaler

An overview of FASTX Analyses is available.

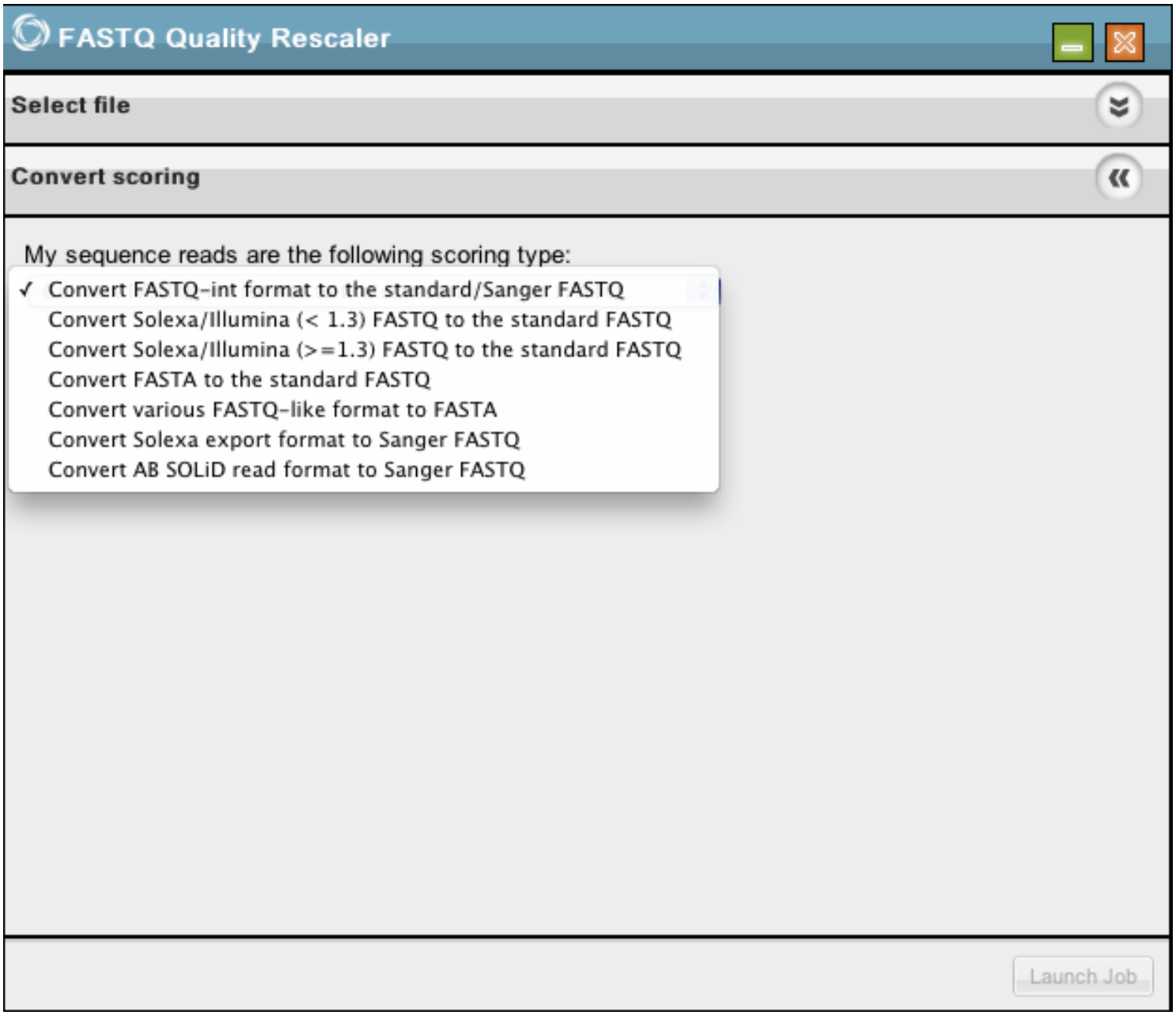
The FASTQ Quality Rescaler updates the base quality scores in your sequence data to use the Phred33 scale adopted by the Sanger Centre and the NCBI Sequence Read Archive. Conversion from Illumina 1.3+ and Solexa is supported.



Select FASTQ Quality Rescaler from within Perform Analyses as described in that section. Click Ok.



Click Browse to select the previously uploaded file you want to convert. Click Convert scoring to continue.

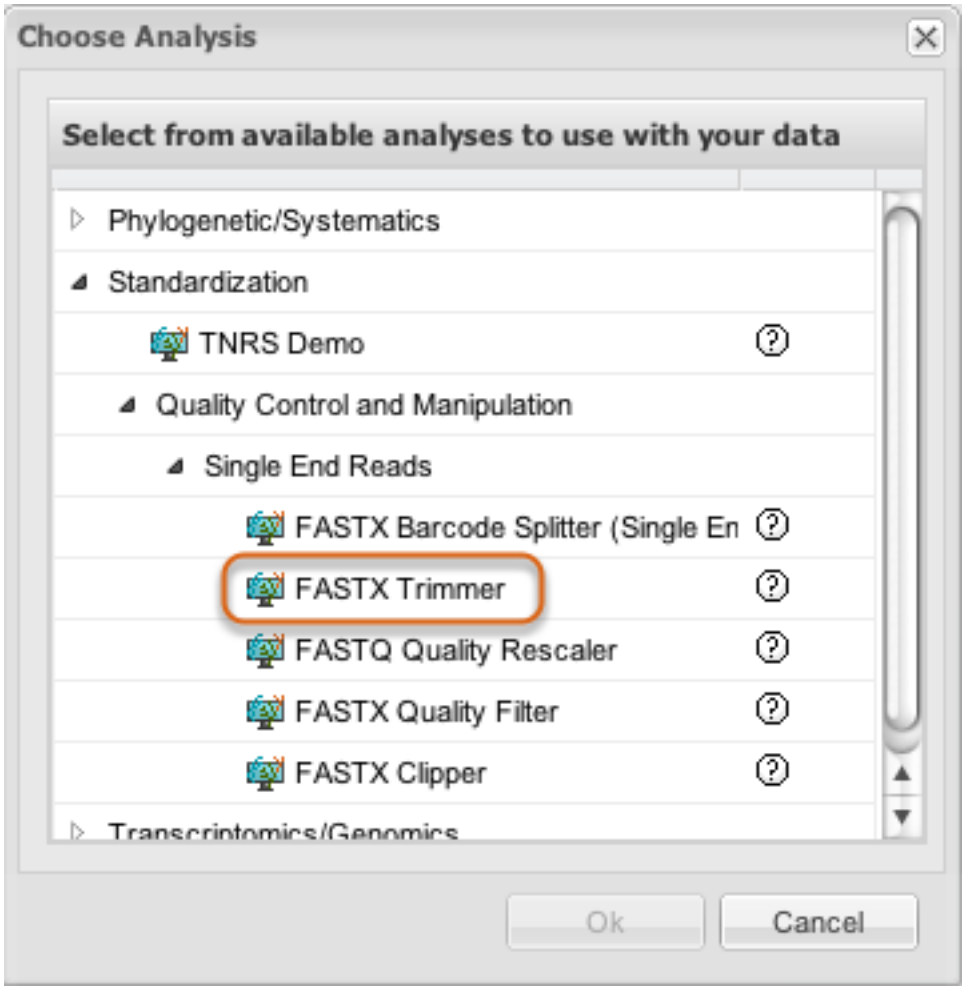


Specify the scoring type used in your read library from the drop-down menu. Click Launch Job.

Enter a name and description for the job and click Ok. See Perform Analyses for information about monitoring the process and where to find your results.

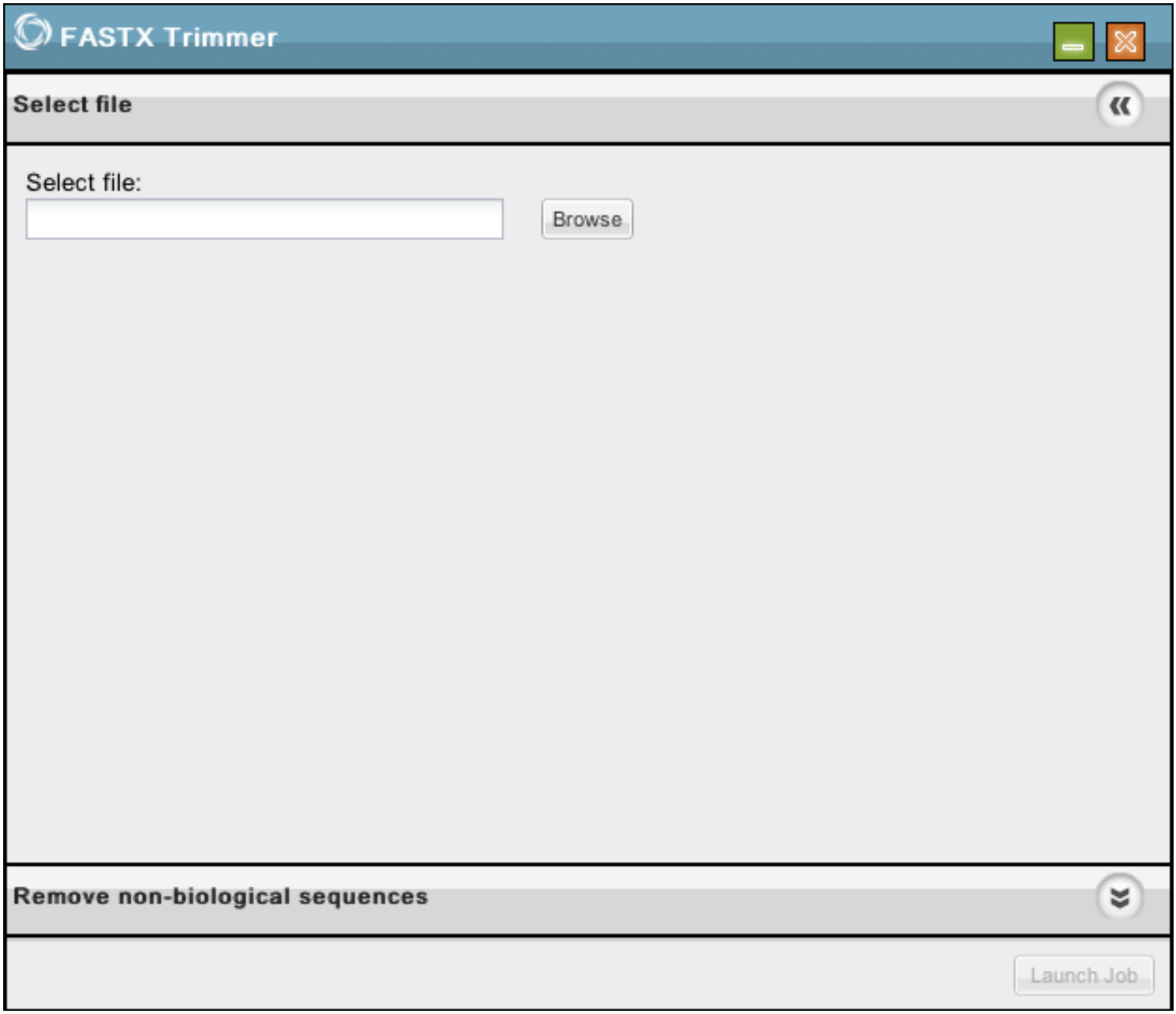
FASTX Trimmer

An overview of FASTX Analyses is available.

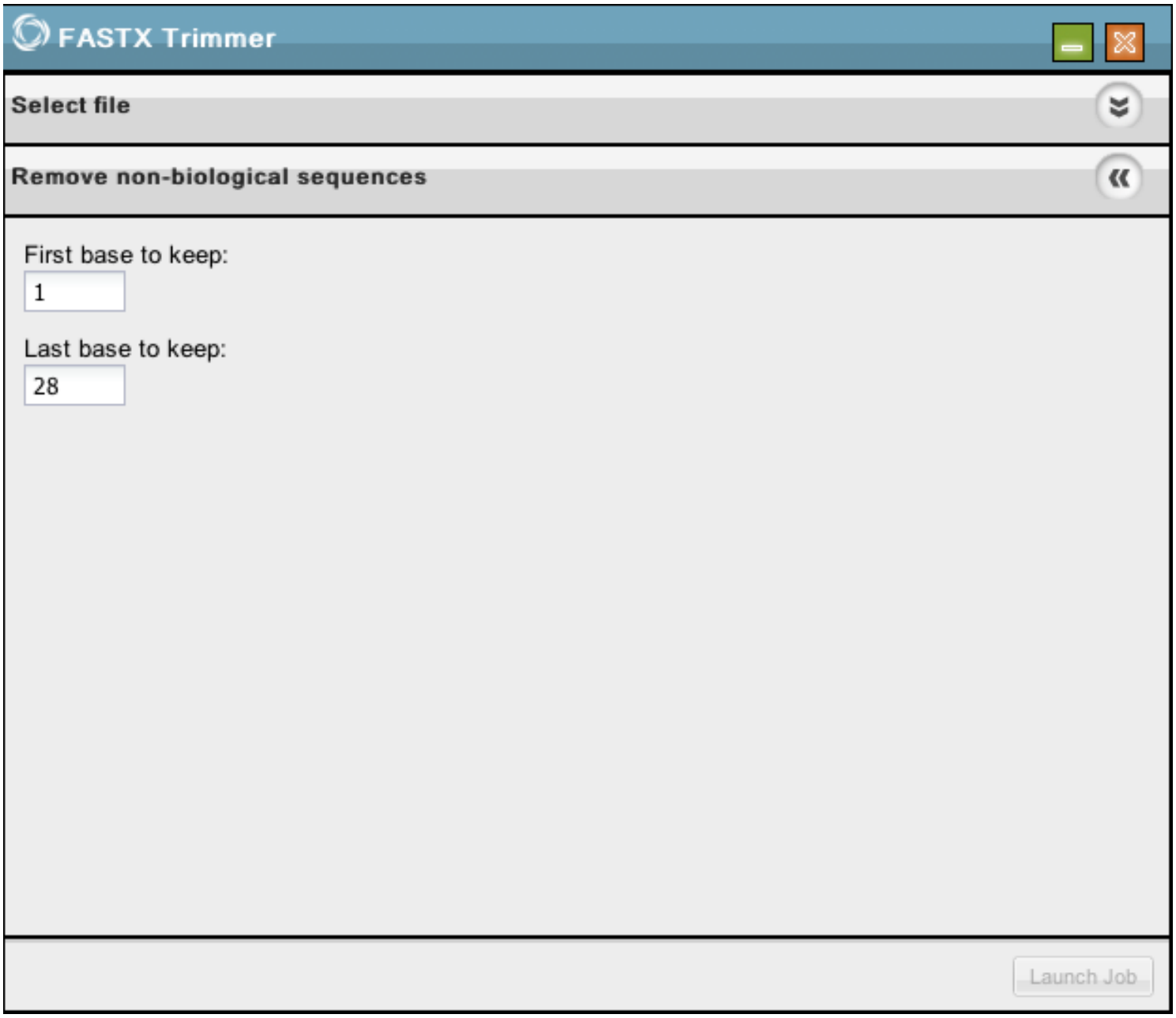


Select FASTX Trimmer from within Perform Analyses as described in that section. Click Ok.

Select file



Click Browse to select your previously uploaded file. Click Remove non-biological sequences.



Keep or modify the default settings. Click Launch Job.

Enter a name and description for the job and click Ok. See Perform Analyses for information about monitoring the process and where to find your results.

Find SNPs Overview

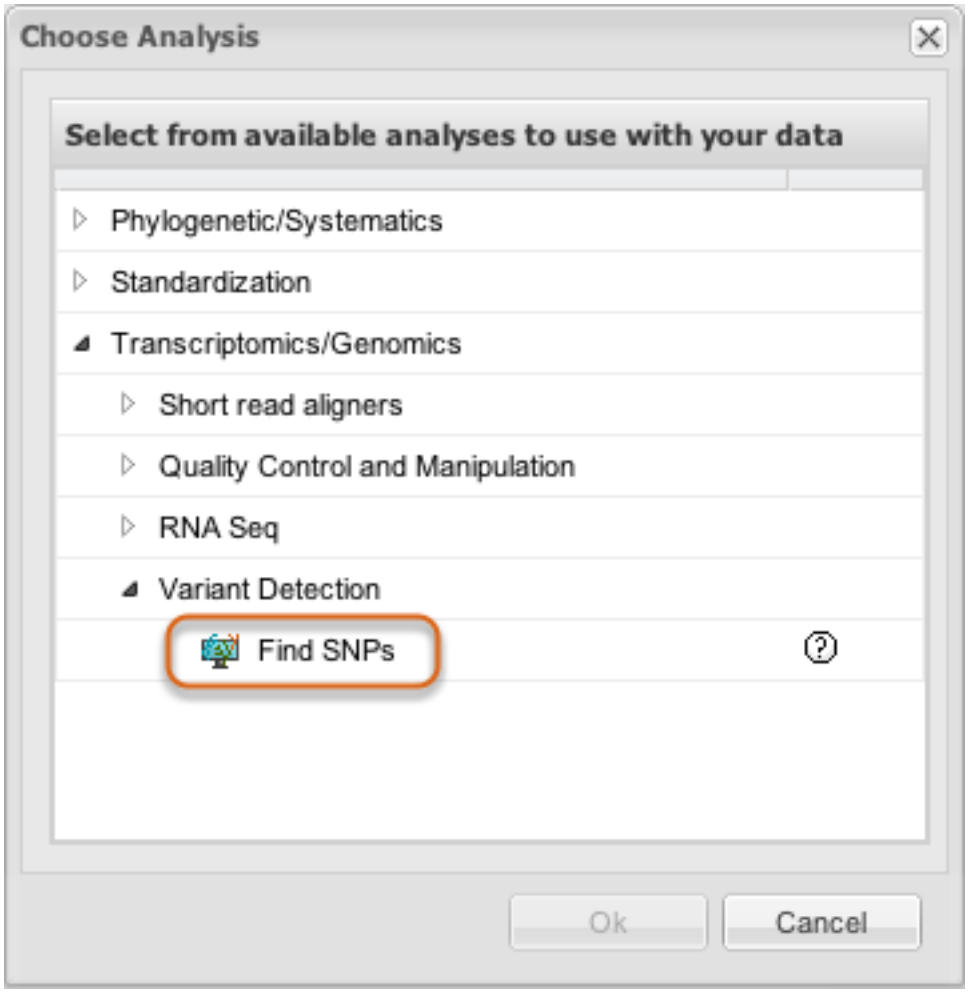
Find SNPs uses SAMtools.

Find SNPs finds variants, or single nucleotide polymorphisms (SNPs), in DNA datasets. You may upload your own existing SAM alignment files that have been derived from one of the supported reference genomes and use them to identify SNPs.

The output of this analysis is a listing of variants in VCF3.3 format.

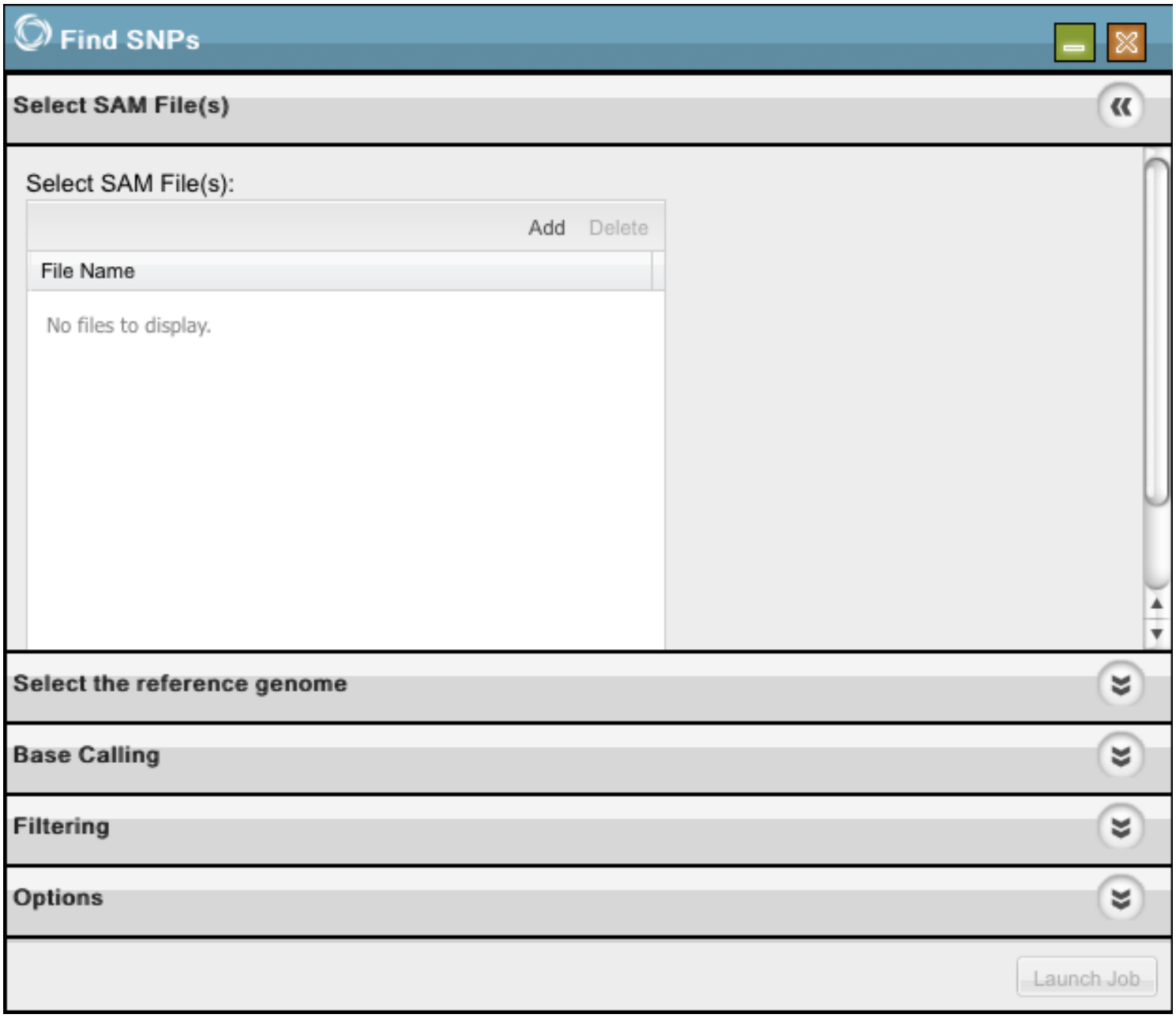
Find SNPs

An overview of Find SNPs is available.



Select Find SNPs from within Perform Analyses as described in that section. Click Ok.

Select SAM File(s)

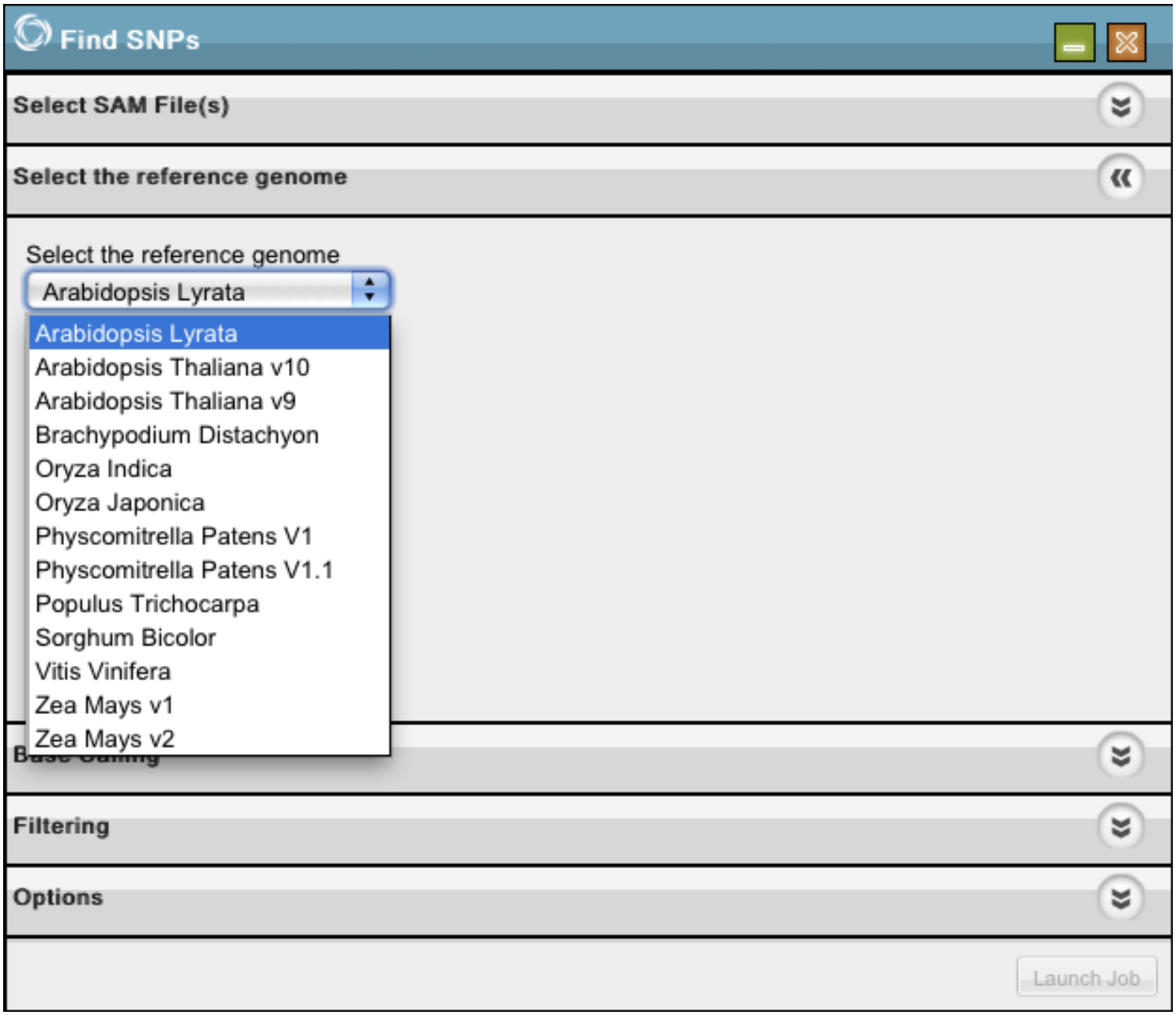


Click Add to choose the previously uploaded SAM files in which you are seeking variants from the reference genome.

There is no limit to the number of files you may select here, but files must be selected one at a time.

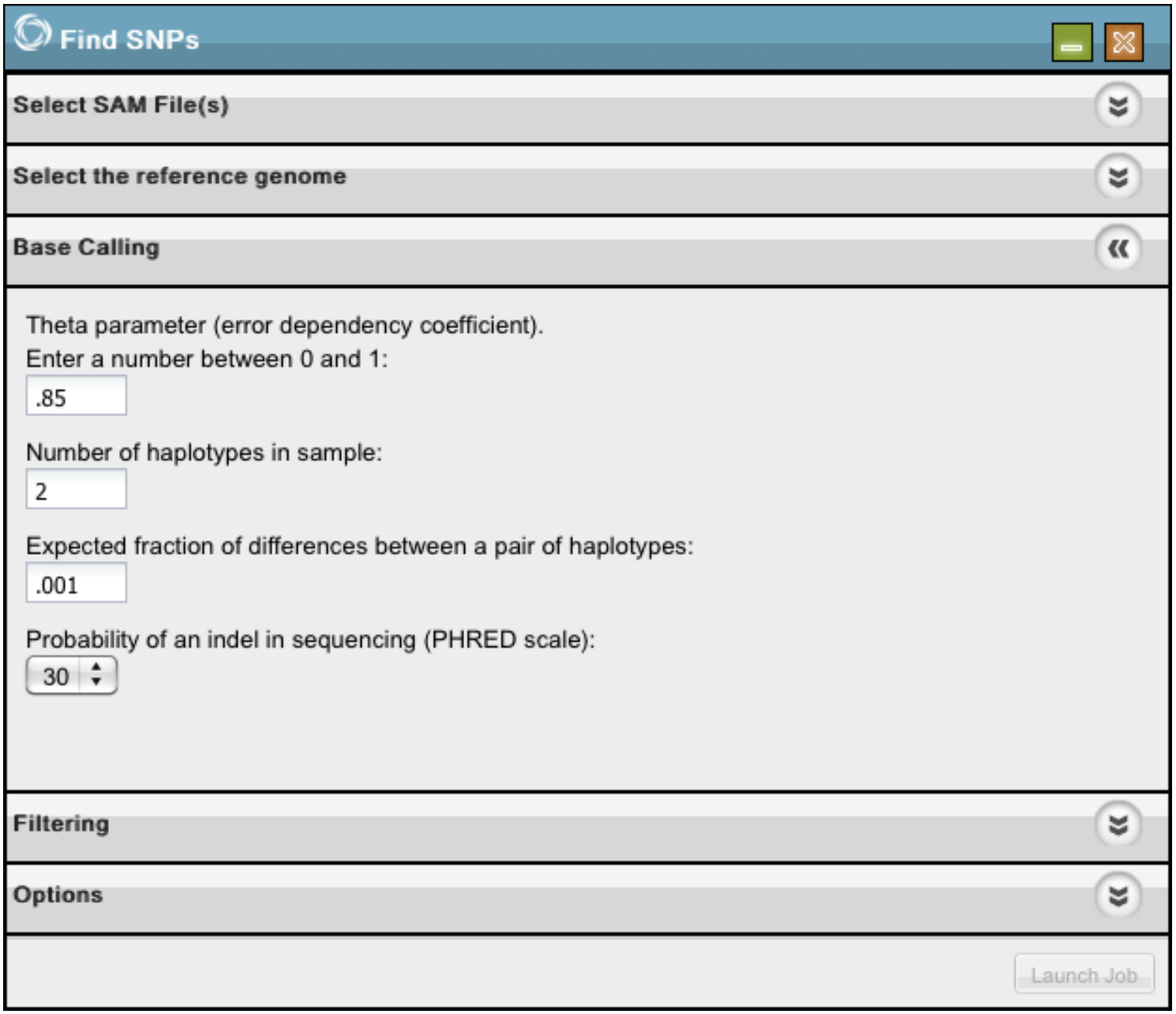
Select a file and Delete will remove files previously selected during this step prior to launching the job.

Select Reference Genome



Select the reference genome to which you will compare your SAM files.

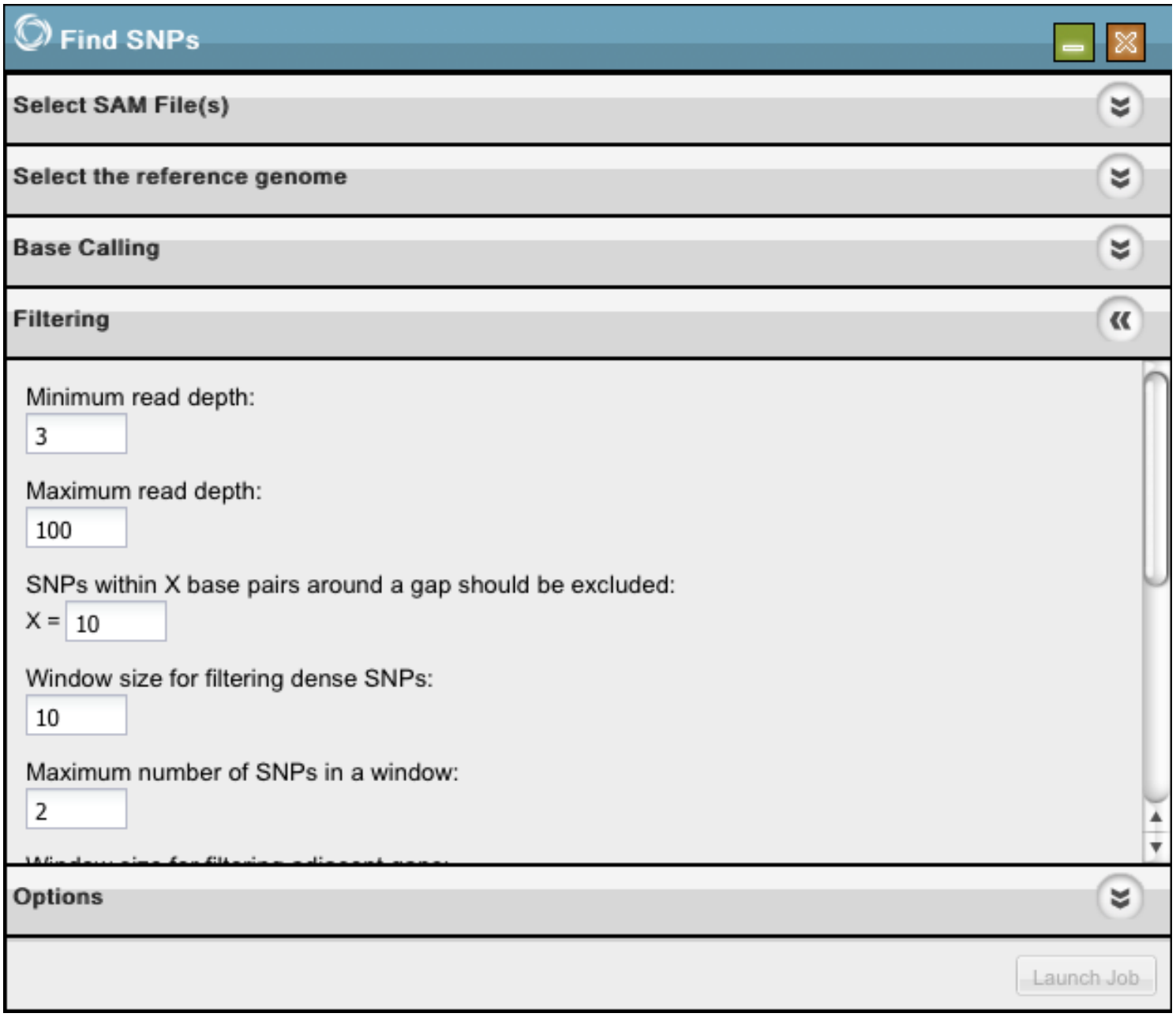
Base Calling



Select the base calling parameters.

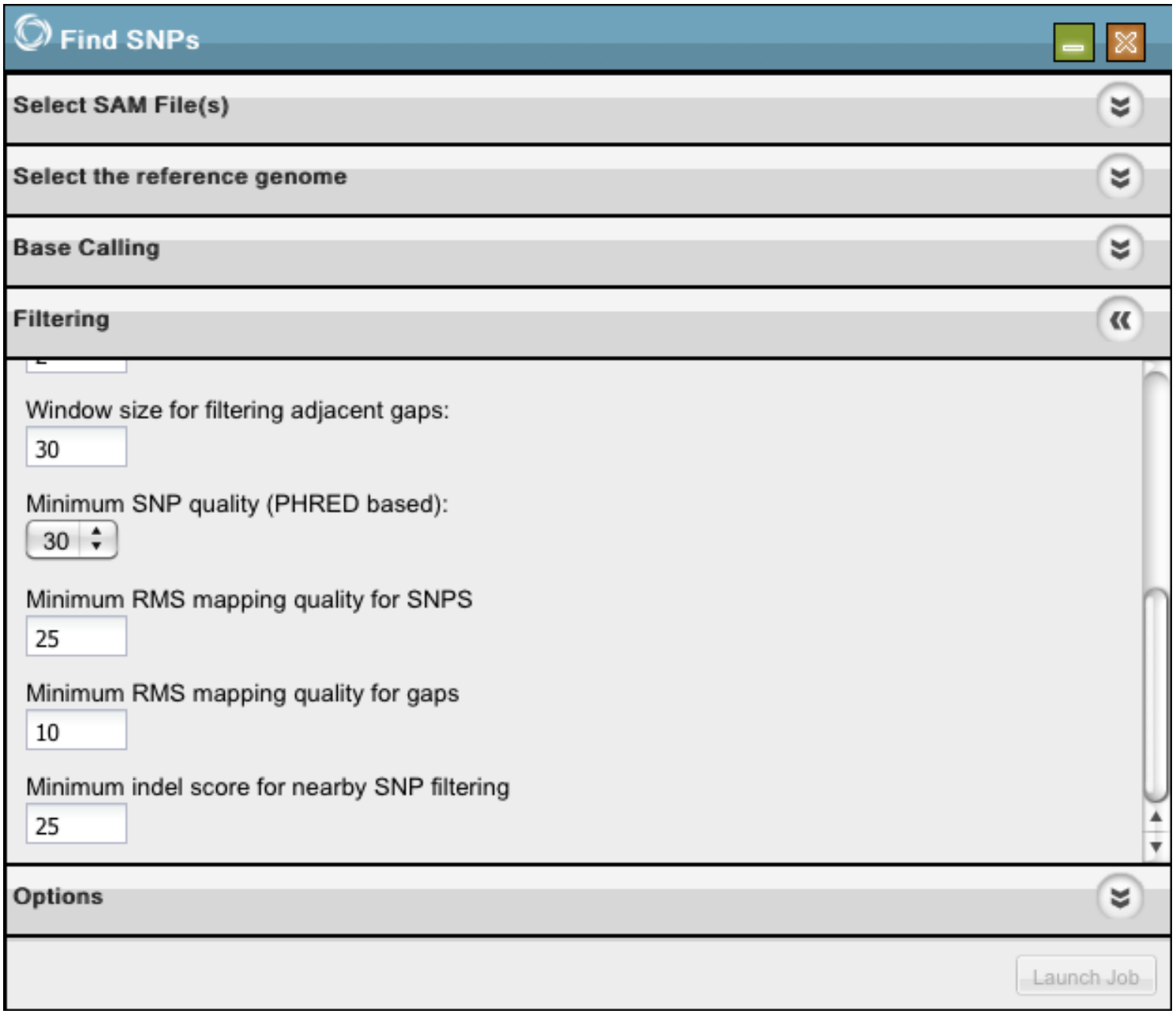
The theta parameter, or error dependency coefficient, uses the maq consensus calling model and defines how much difference will be tolerated when calculating variance, assuming these differences to be natural fluctuations or other error rather than different sequences. For more details on these parameter settings, please see SAMtools and Maq.

Filtering (part one)

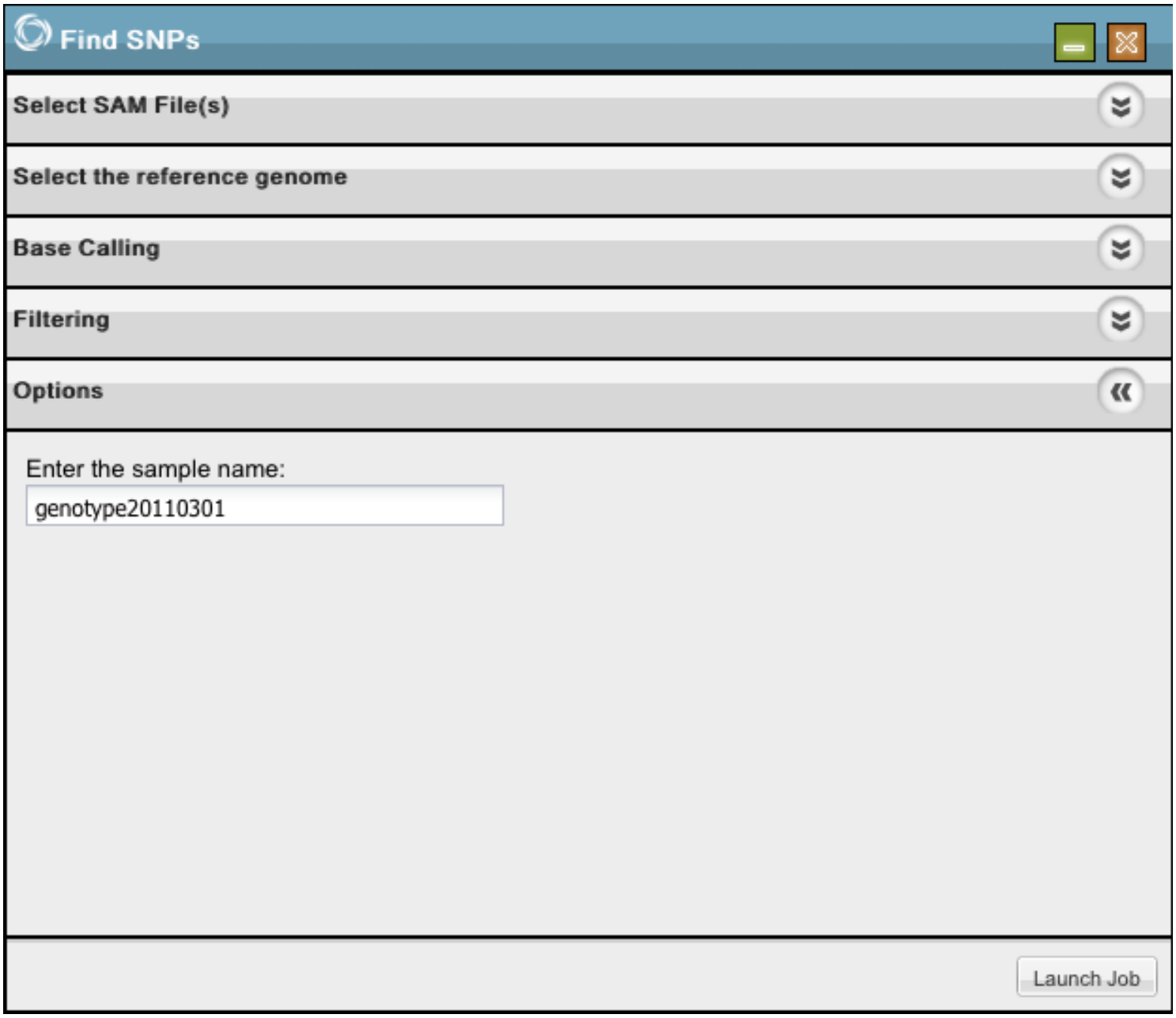


Enter your desired filtering parameters, here and below. For details on the filtering parameters, please see SAMtools.

Filtering (part two)



Options



Enter a name for your genotype sample to make it easier for you to keep track of multiple VCF data records. Click Launch Job.

Enter a name and description for the job and click Ok. See Perform Analyses for information about monitoring the process and where to find your results.

Independent Contrasts Overview

Phylogenetic Independent Contrasts (PIC) is a subset of phylogenetic comparative methods, which use information on the evolutionary relationships of organisms (phylogenetic trees) to test for correlated evolutionary changes in two or more traits. PIC is a statistically-based approach that uses the phylogenetic tree and evolutionary branch lengths as a guide to determine whether two or more quantitative characters are evolutionarily correlated. PIC can help users discern between characters that are similar because of a common evolutionary history from those which are similar for other reasons, such as an adaptive response to environmental conditions.

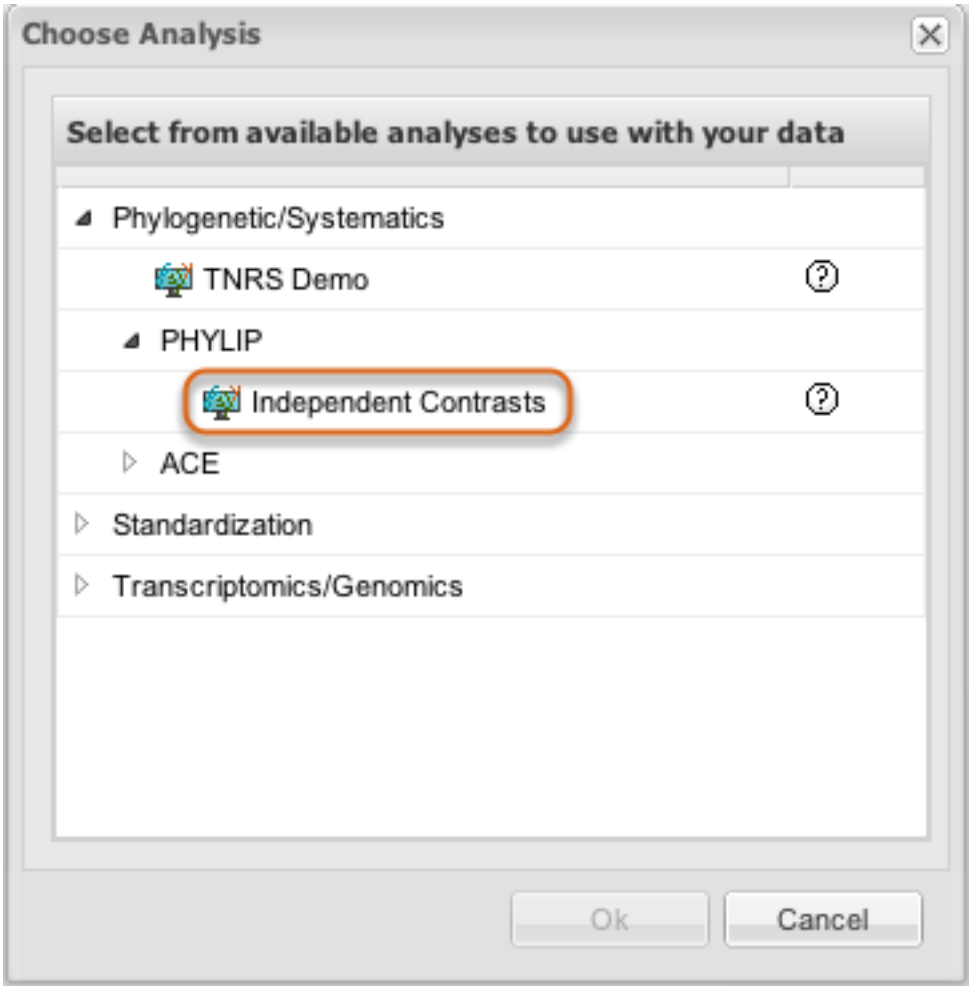
For someone doing data analysis, PIC can be considered as a new set of characters with evolution history subtracted. Thus the correlation between two or more PIC characters becomes meaningful.

PIC uses the Contrast program from PHYLIP.

This method originated in this paper: Felsenstein, J. 1985. Phylogenies and the comparative method. American Naturalist 125:1-15.

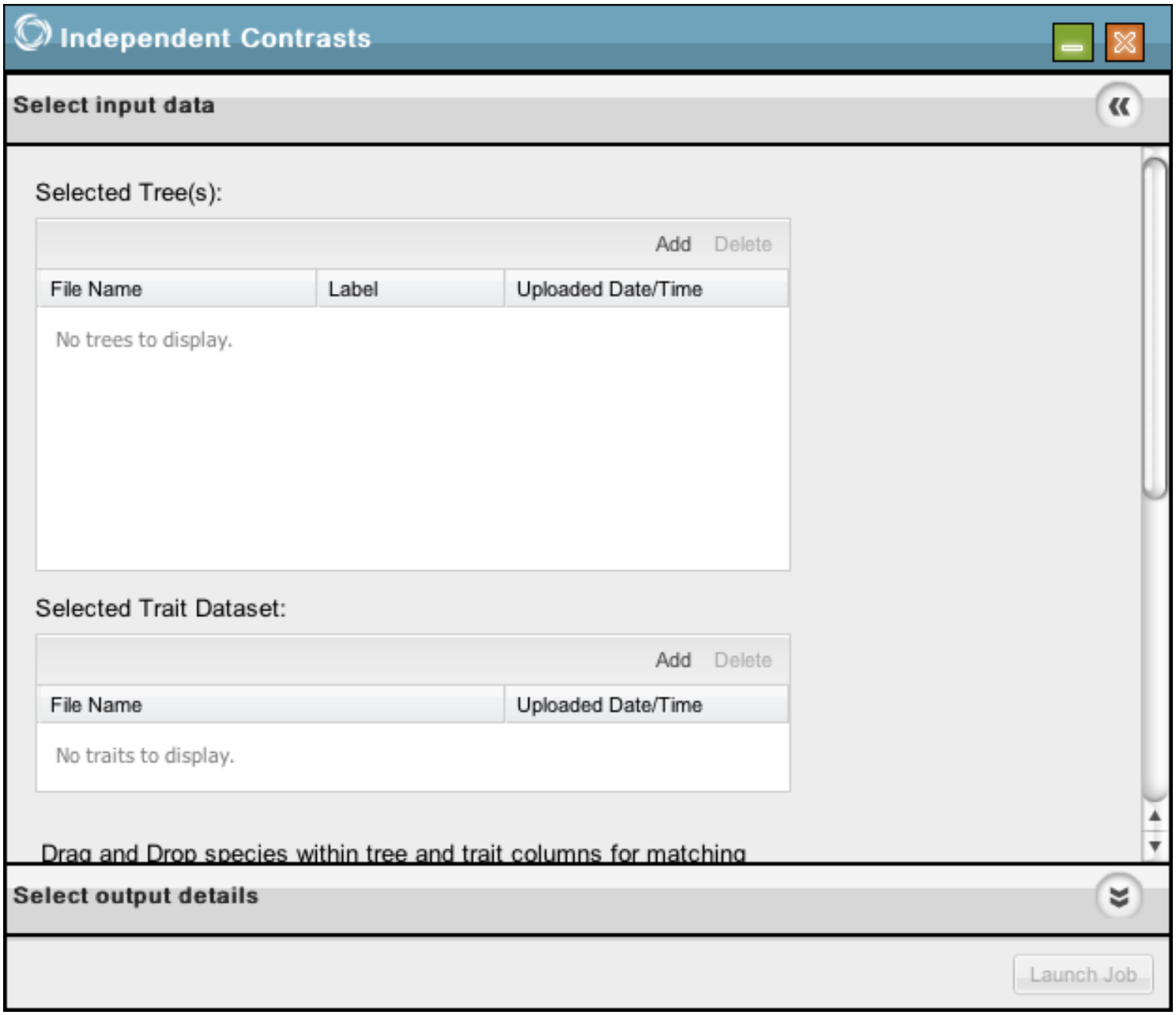
Independent Contrasts

An overview of Independent Contrasts is available.



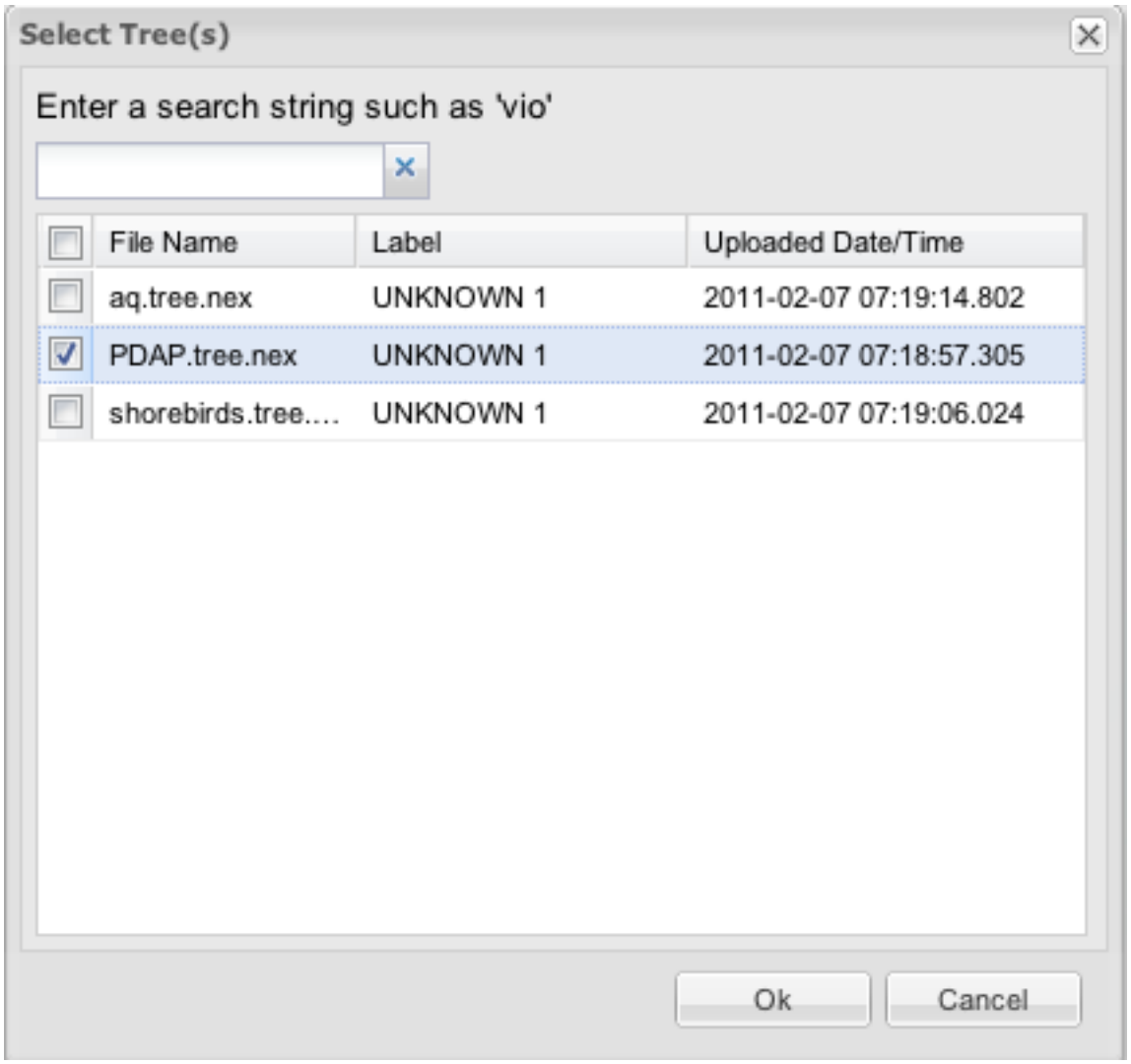
Select Independent Contrasts from within Perform Analyses as described in that section. Click Ok.

Select input data



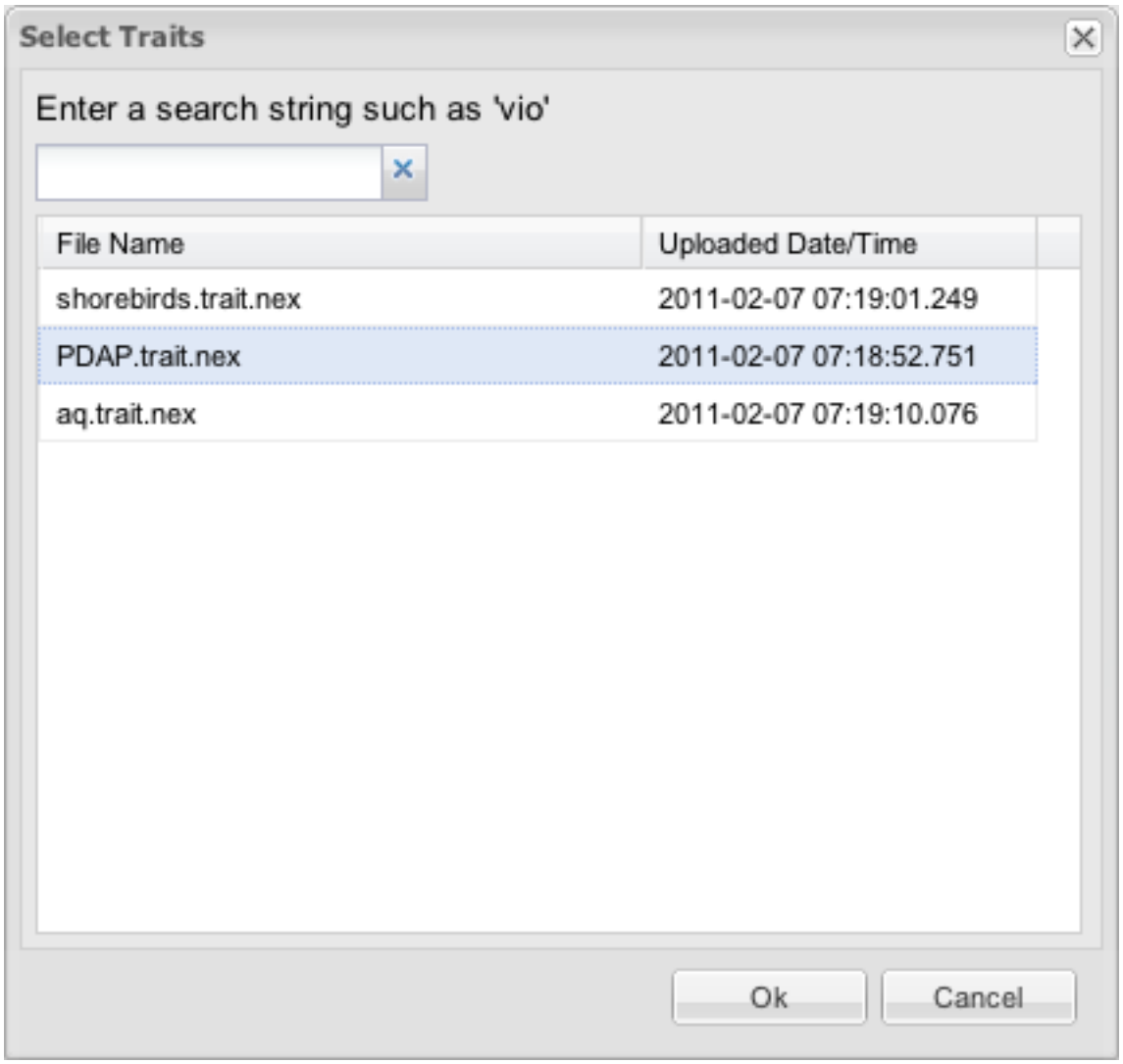
Data needs to be uploaded to the Discovery Environment in advance. Click Add in Selected Tree(s) and Selected Trait Dataset to choose appropriate tree and trait files from the boxes shown next.

Select Tree or Trees



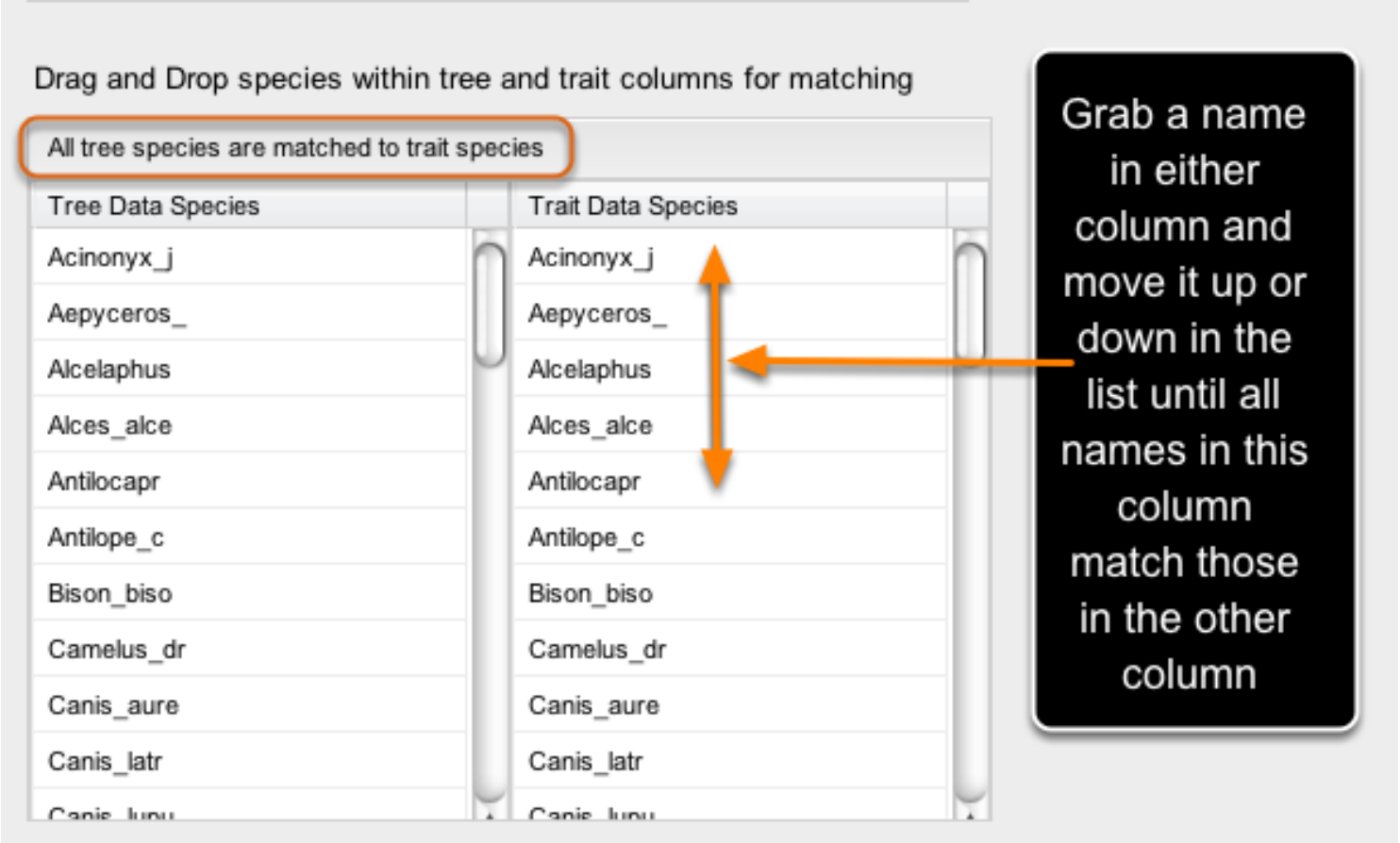
Highlight your desired file(s) and click Ok.

Select Traits



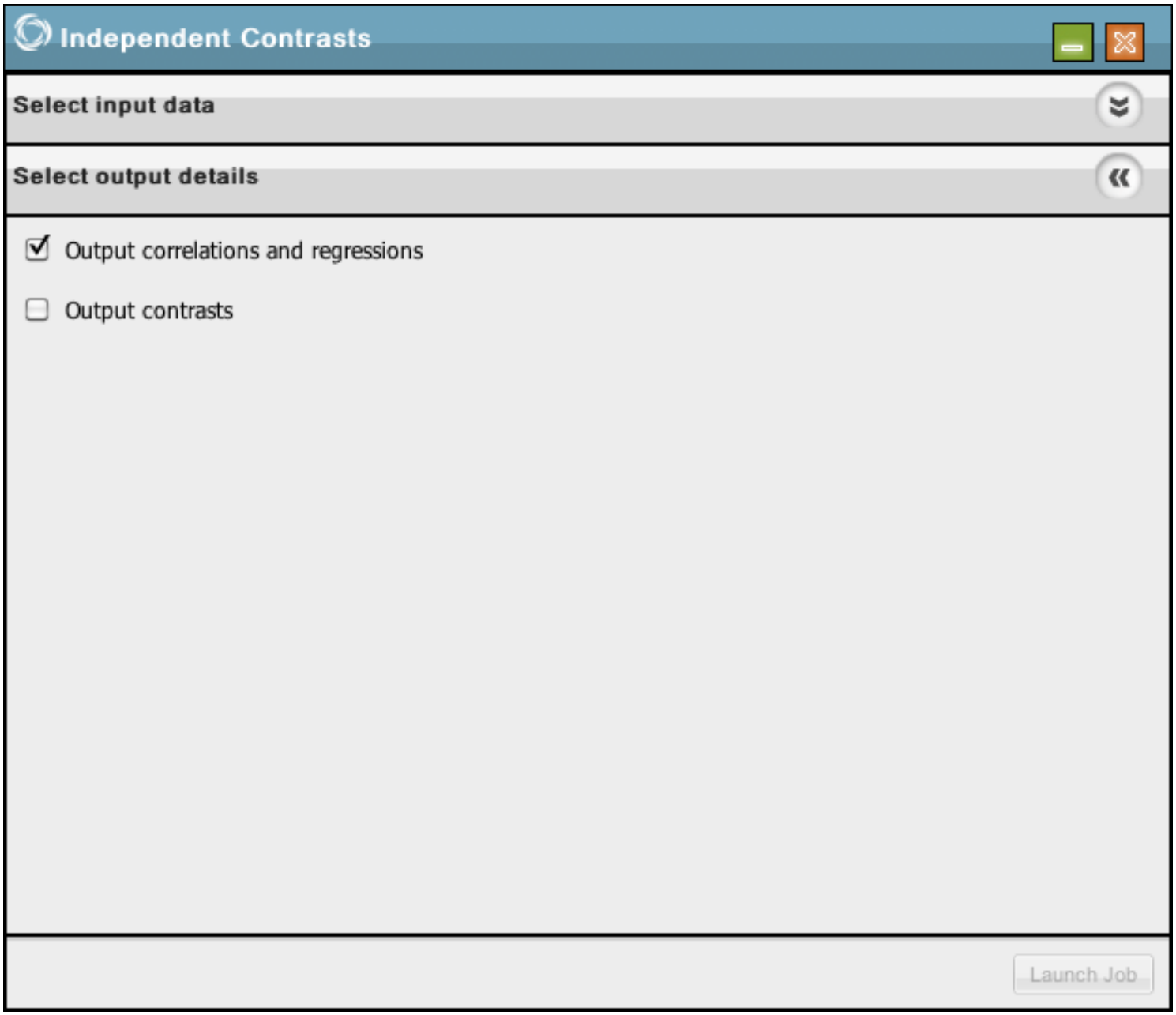
Highlight your desired file and click Ok.

Match Data


Drag and Drop species within tree and trait columns for matching
All tree species are matched to trait species
Tree Data Species
Trait Data Species
Acinonyx_j
Aepyceros_
Alcelaphus
Alces_alce
Antilocapr
Antilope_c
Bison_biso
Camelus_dr
Canis_aure
Canis_latr
Grab a name in either column and move it up or down in the list until all names in this column match those in the other column

Hold the left mouse button to drag and swap to move species data up and down until all tree species and trait species are matched. When the text above the table shows All tree species are matched to trait species, click Select output details.

Select Output Details



Next, click Select output details. You can select Output correlations and regressions, Output contrasts or both if desired. Neither is required.

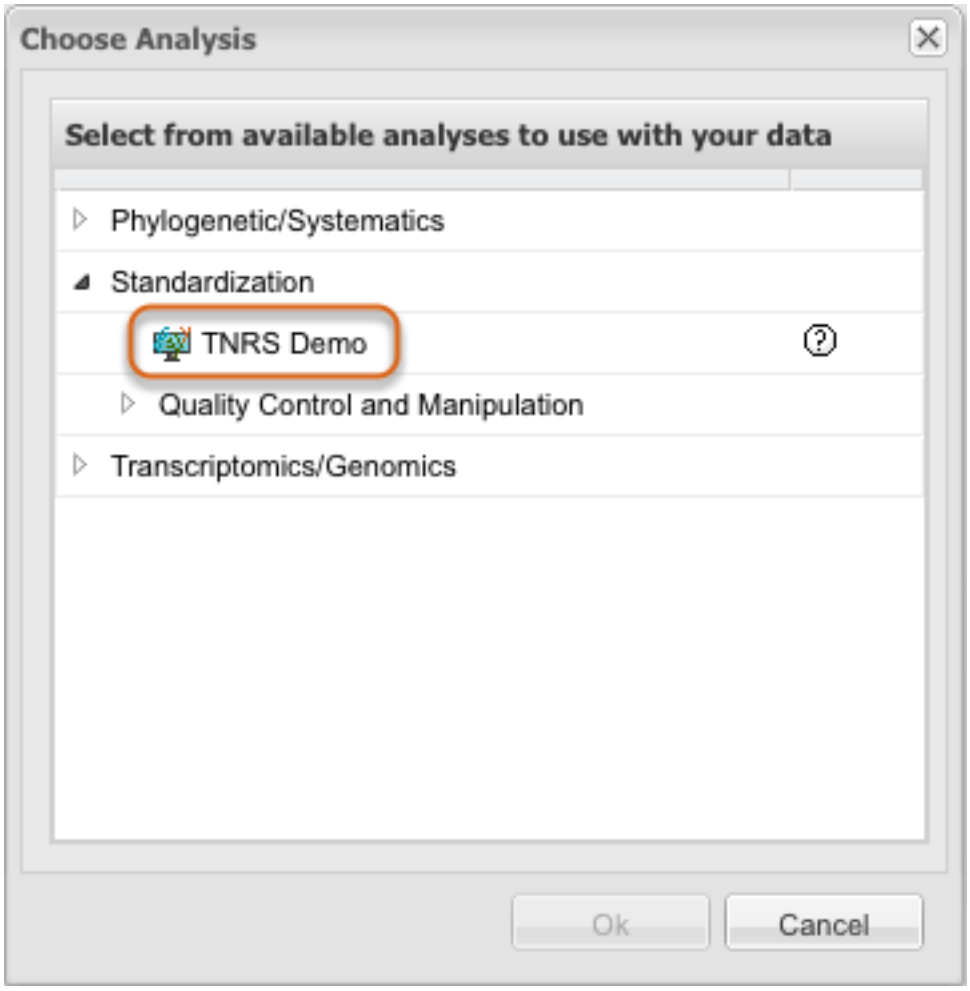
Click Launch Job. See Perform Analyses for information about monitoring the process and where to find your results.

Taxonomic Name Resolution Service (TNRS) Demo

Accepts a list of taxa and checks them against a database of canonical names to return both exact and possible matches. Uses exact (via database queries) and fuzzy matching (via Taxamatch) to compare a list of submitted names with a standardized database.

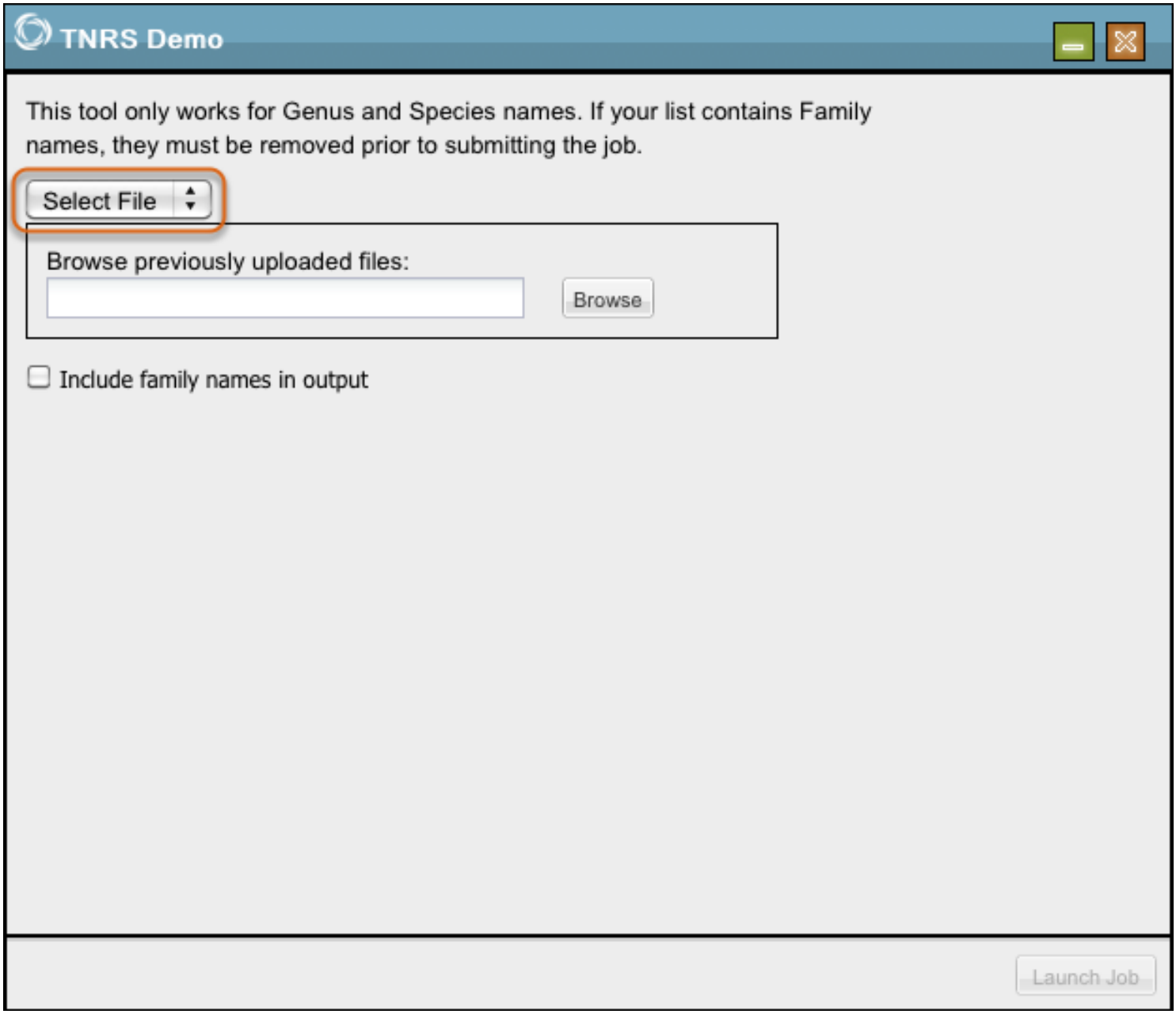
Author data and further information available at: http://tnrs.iplantcollaborative.org.

The tool was identified for inclusion by the iPlant Tree of Life working group.



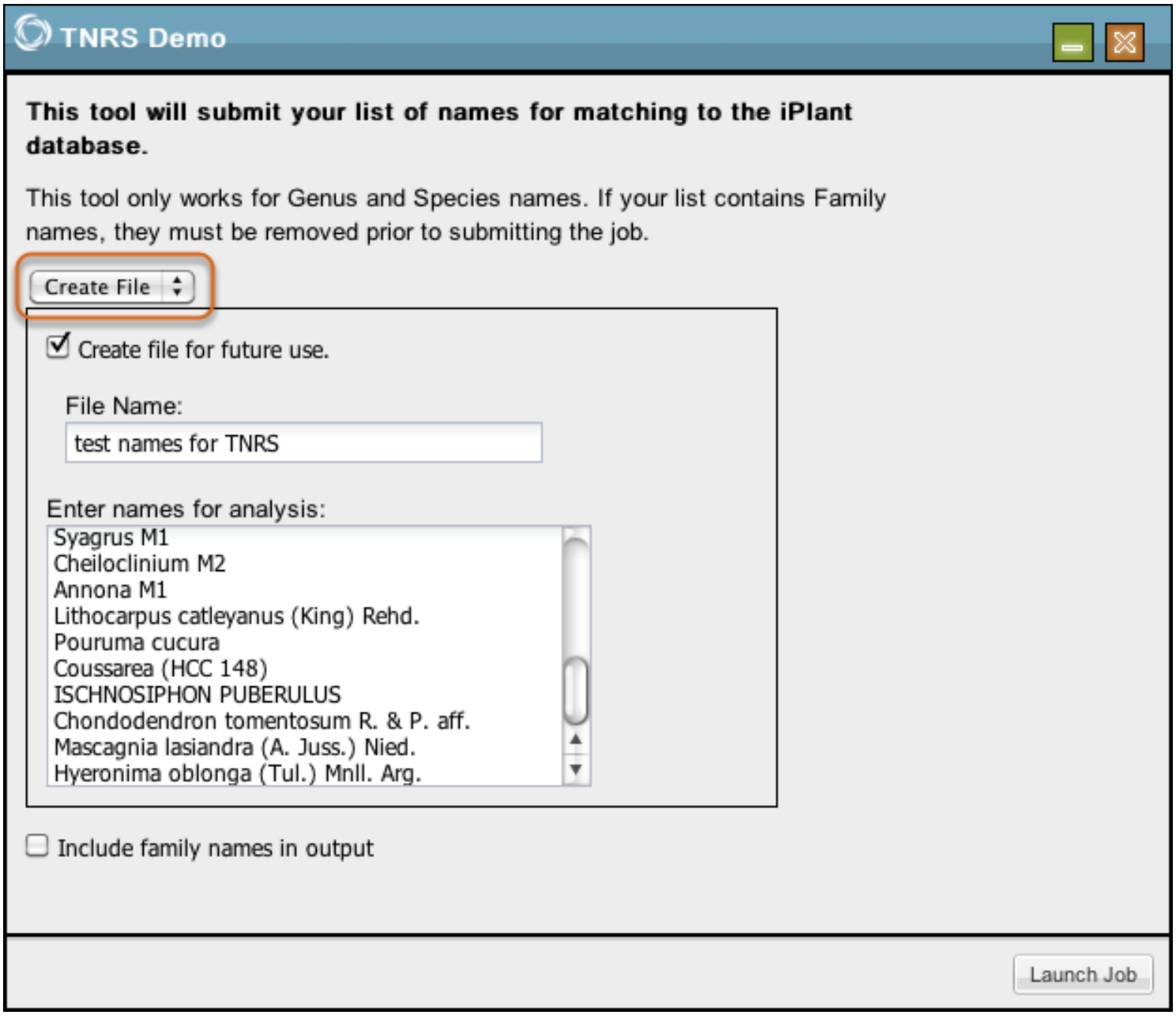
Select TNRS Demo from within Perform Analyses as described in that section. Click Ok.

Submit a list of names


TNRS Demo
This tool only works for Genus and Species names. If your list contains Family names, they must be removed prior to submitting the job.
Select File
Browse previously uploaded files:
Browse
Include family names in output
Launch Job

You may submit a previously uploaded list of names by selecting Select File from the drop-down menu. Click Launch Job.

Enter a list of names

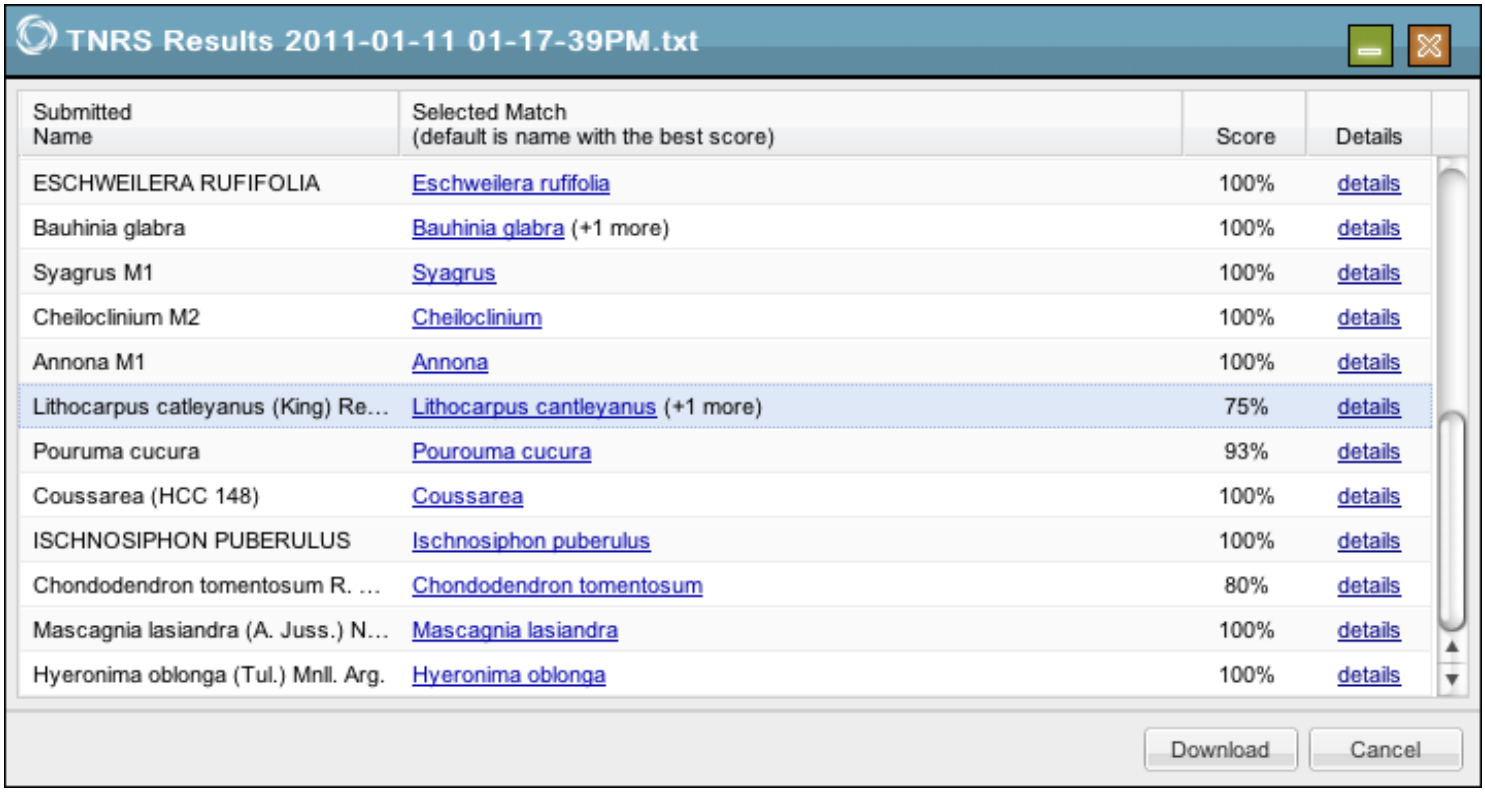

TNRS Demo
This tool will submit your list of names for matching to the iPlant database.
This tool only works for Genus and Species names. If your list contains Family names, they must be removed prior to submitting the job.
Create File
Create file for future use.
File Name:
test names for TNRS
Enter names for analysis:
Syagrus M1
Cheiloclinium M2
Annona M1
Lithocarpus catleyanus (King) Rehd.
Pouruma cucura
Coussarea (HCC 148)
ISCHNOSIPHON PUBERULUS
Chondodendron tomentosum R. & P. aff.
Mascagnia lasiandra (A. Juss.) Nied.
Hyeronima oblonga (Tul.) Mnll. Arg.
Include family names in output
Launch Job

You may enter a list of names directly into the tool by selecting Create File from the drop-down menu.

If you check the box next to Create file for future use, you can then enter a file name and the file will be available to you in Manage Data. Click Launch Job.

Enter a name and description for the job and click Ok. See Perform Analyses for information about monitoring the process and where to find your results.

View your results

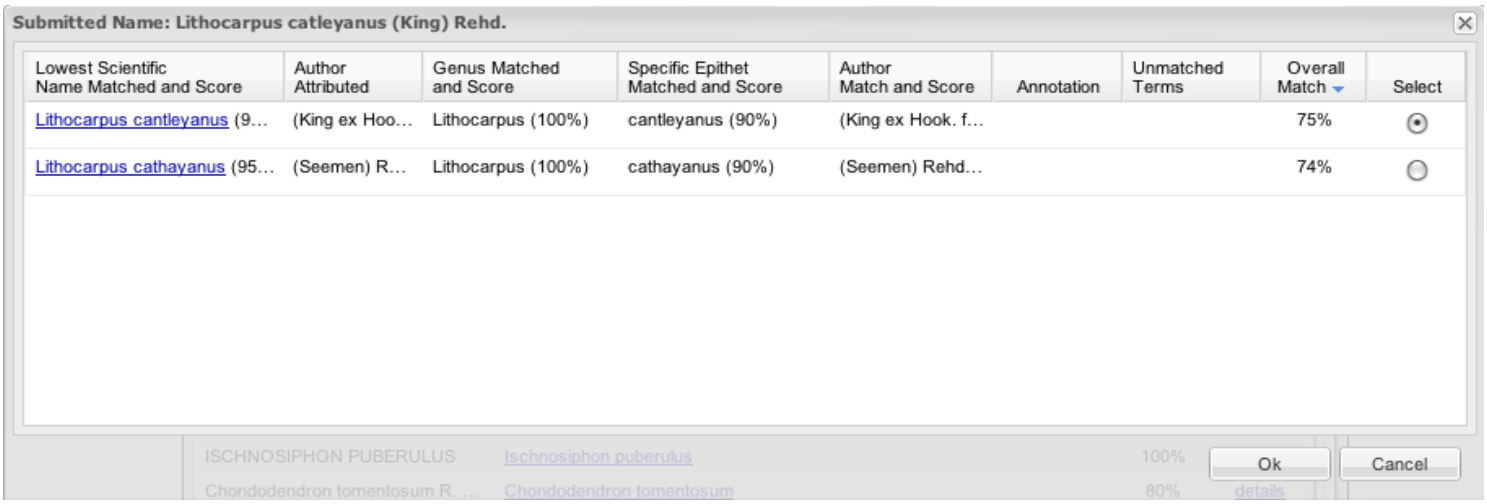


Submitted Name	Selected Match (default is name with the best score)	Score	Details
ESCHWEILERA RUFIFOLIA	Eschweilera rufifolia	100%	details
Bauhinia glabra	Bauhinia glabra (+1 more)	100%	details
Syagrus M1	Syagrus	100%	details
Cheiloclinium M2	Cheiloclinium	100%	details
Annona M1	Annona	100%	details
Lithocarpus catleyanus (King) Re...	Lithocarpus cantleyanus (+1 more)	75%	details
Pouruma cucura	Pourouma cucura	93%	details
Coussarea (HCC 148)	Coussarea	100%	details
ISCHNOSIPHON PUBERULUS	Ischnosiphon puberulus	100%	details
Chondodendron tomentosum R. ...	Chondodendron tomentosum	80%	details
Mascagnia lasiandra (A. Juss.) N...	Mascagnia lasiandra	100%	details
Hyeronima oblonga (Tul.) Mnll. Arg.	Hyeronima oblonga	100%	details

Click the name of a Selected Match to view the database entry for the item on TROPICOS. Matches are given a percent score based on the probability of the match. Further details are available by clicking details.

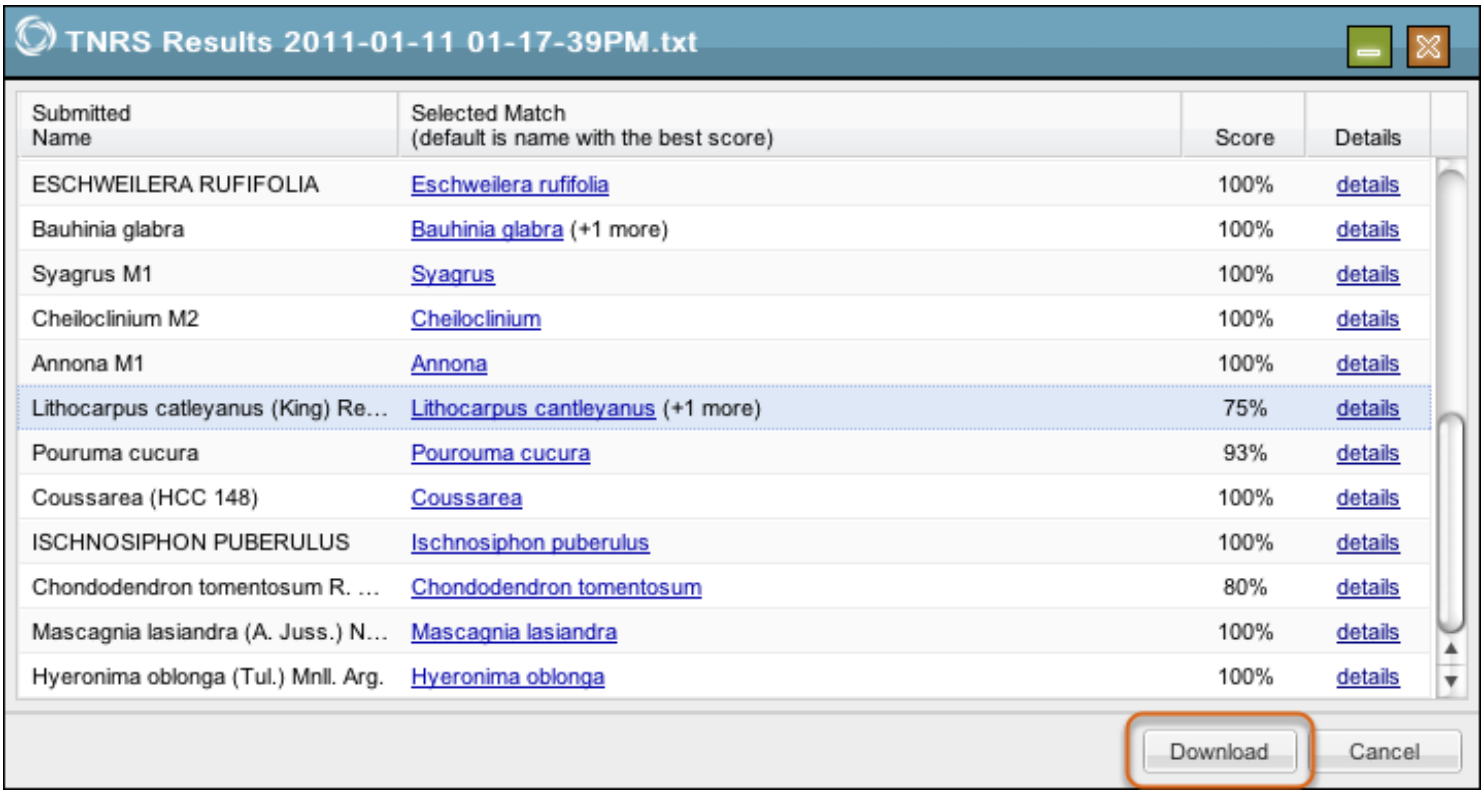
When more than one item is found as a possible match, this is noted. Click details to view more details about possible matches found to determine which match is best.

Choose from among possible matches



When more than one item is found as a possible match, you may view details in the TROPICOS database by clicking each matched name. Denote which one you want to appear in your final list by placing a mark in the circle to the right. Click Ok.

Download results

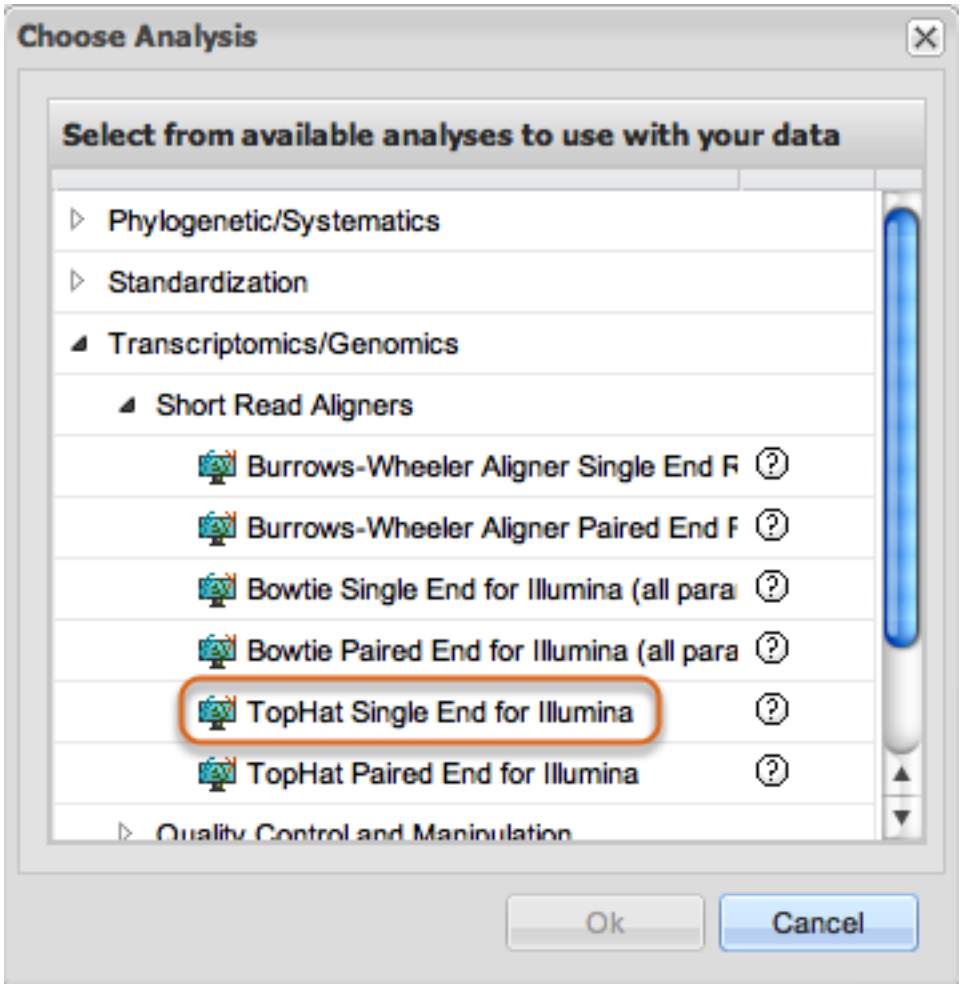


When the main results list shows the names you want to accept, click Download to download a .csv file of your results.

Note that when no author was entered, no authority returned indicates a case when there are multiple records having the same scientific name but different authorities. Each item listed in this instance is a synonym. A future release will add support to return the authority for the accepted name even when no author is entered as well as the ability to match from family to variety.

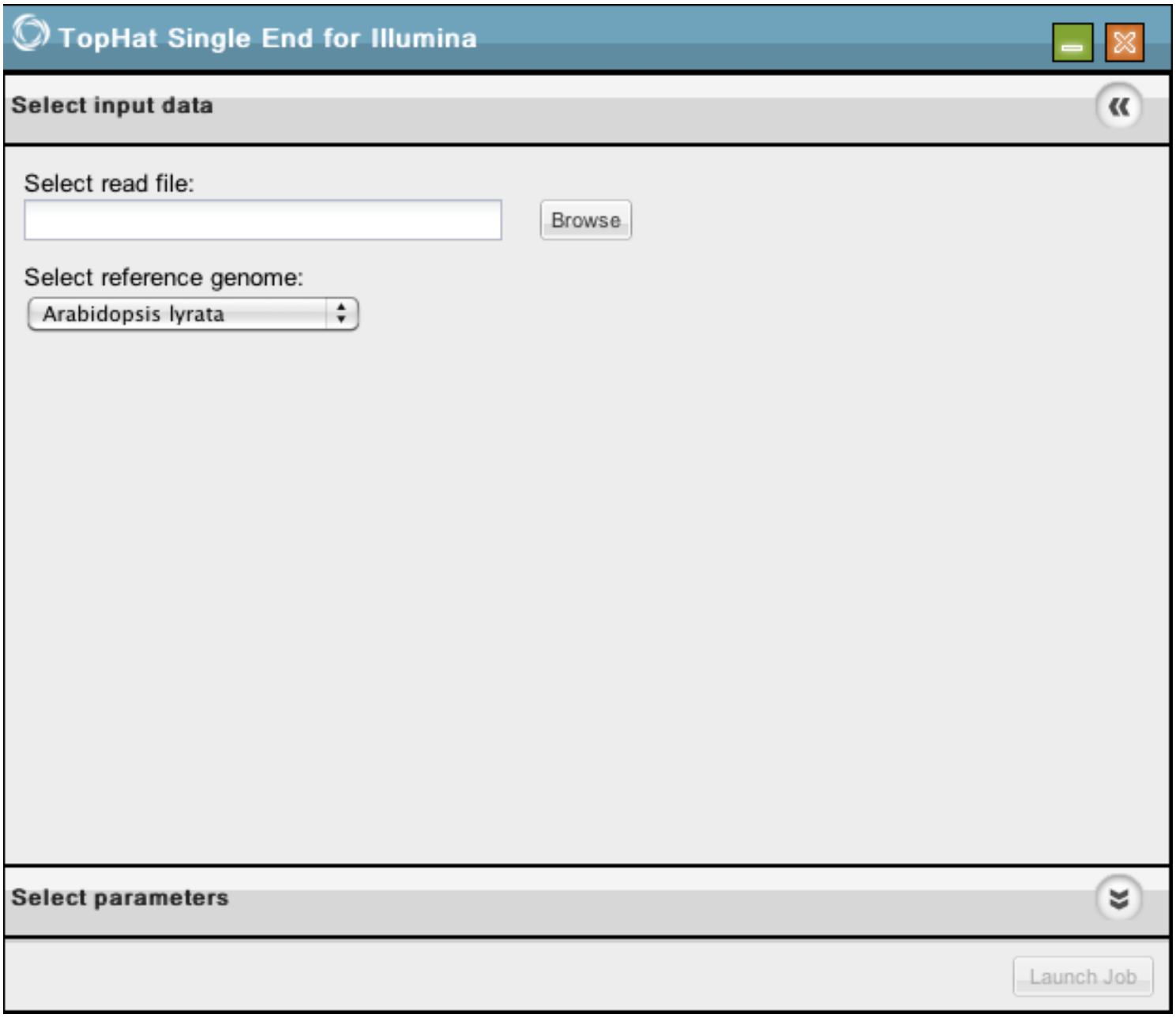
TopHat Single End for Illumina

This analysis uses TopHat. The configuration options are set to be optimal for single end reads derived from Illumina sequencing technology, not 454, ABI, or PacBio. A similar analysis is available for paired end reads.



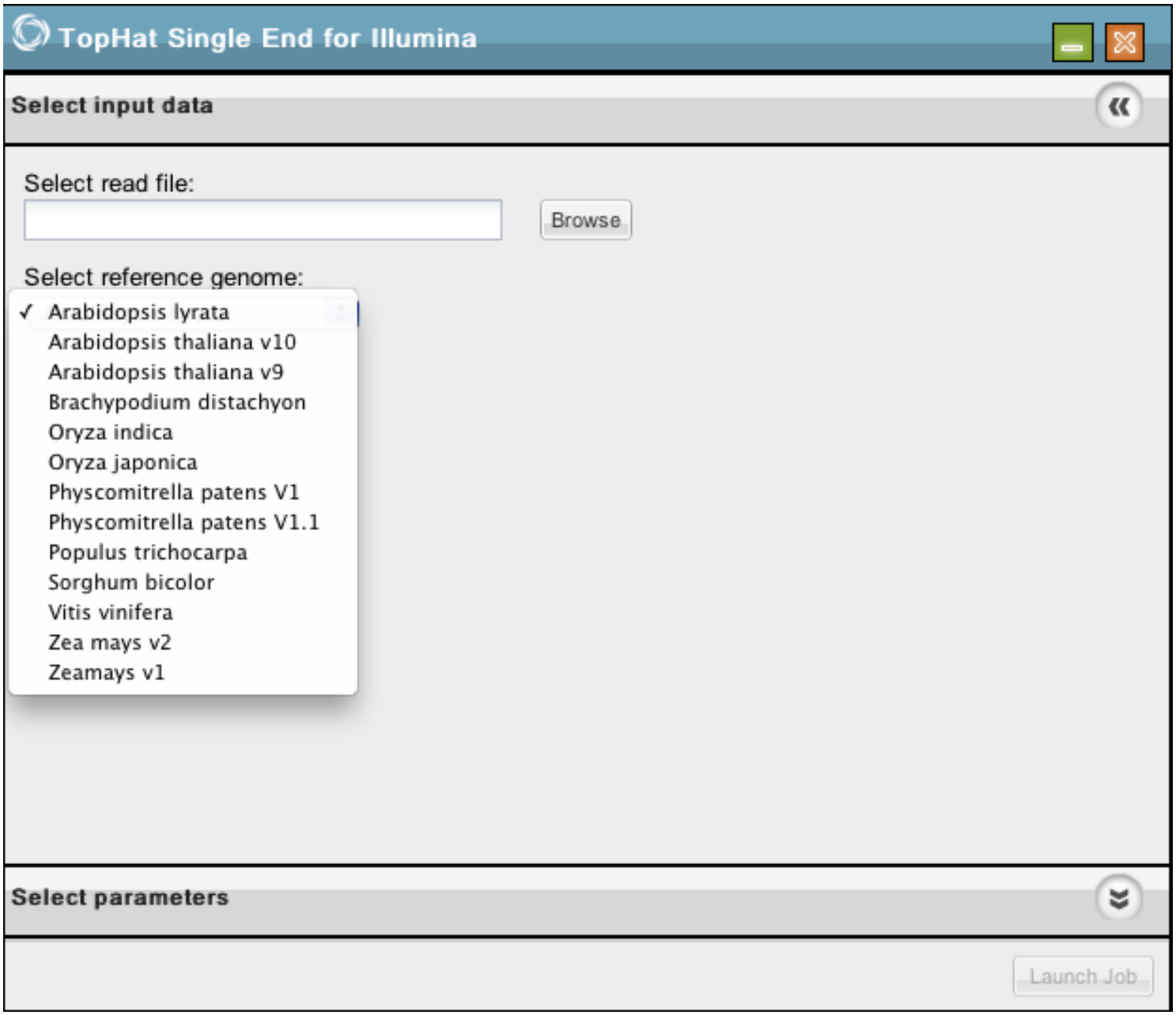
Select TopHat Single End for Illumina from within Perform Analyses as described in that section. Click Ok.

Select input data



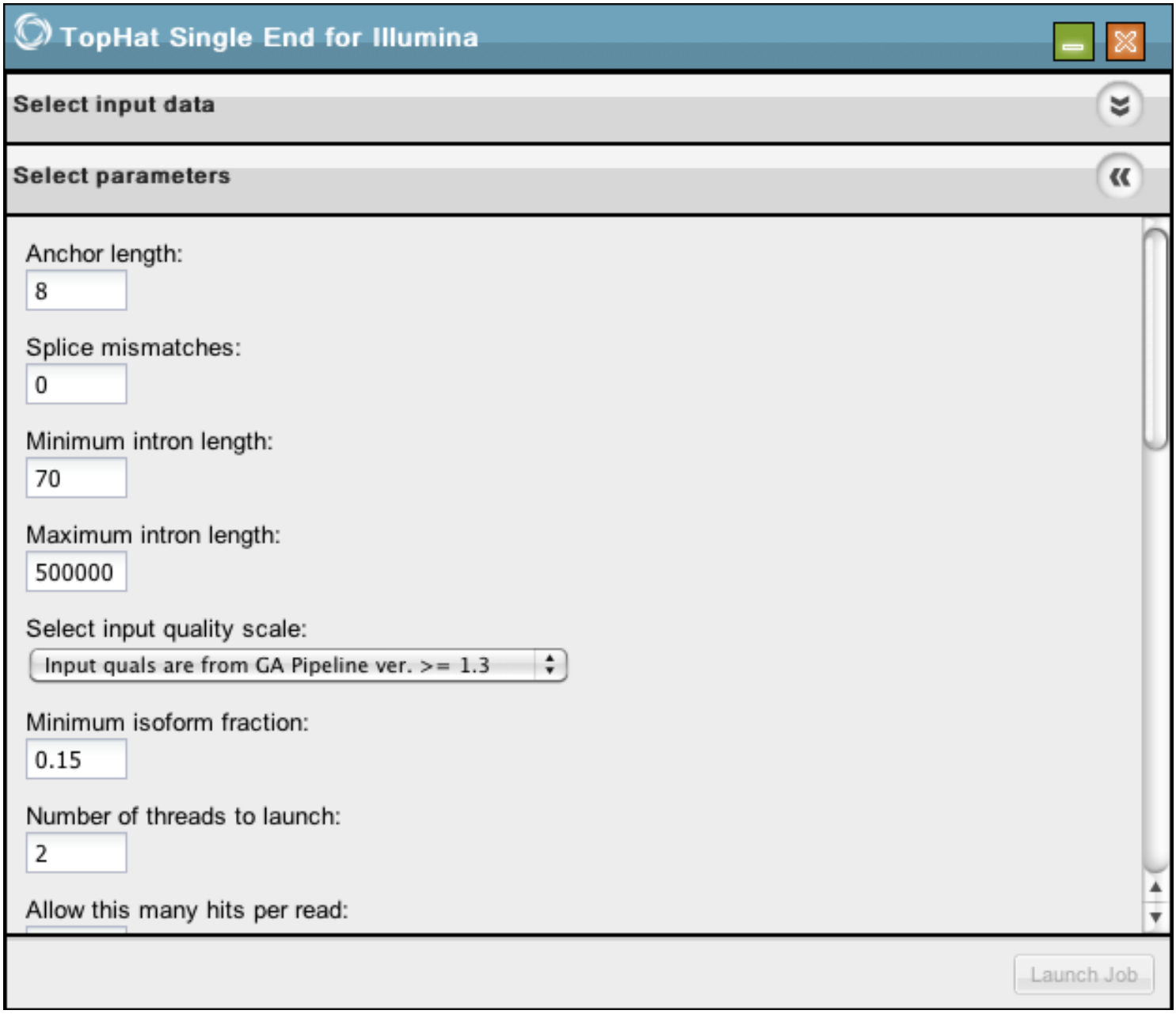
Click Browse to choose the previously uploaded read file you wish to align to a reference genome.

Select Reference Genome



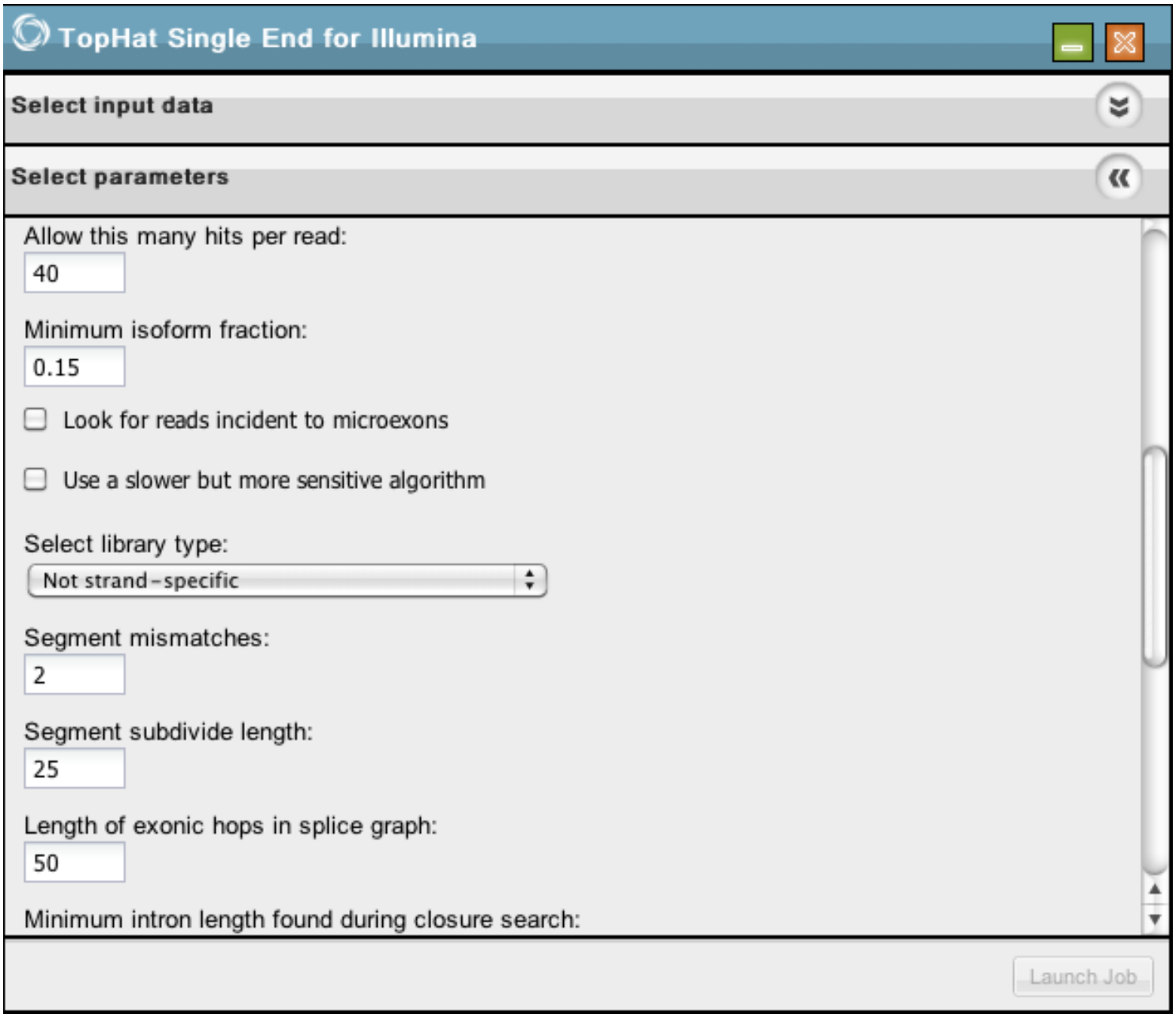
Select the reference genome.

Select Parameters (part one)

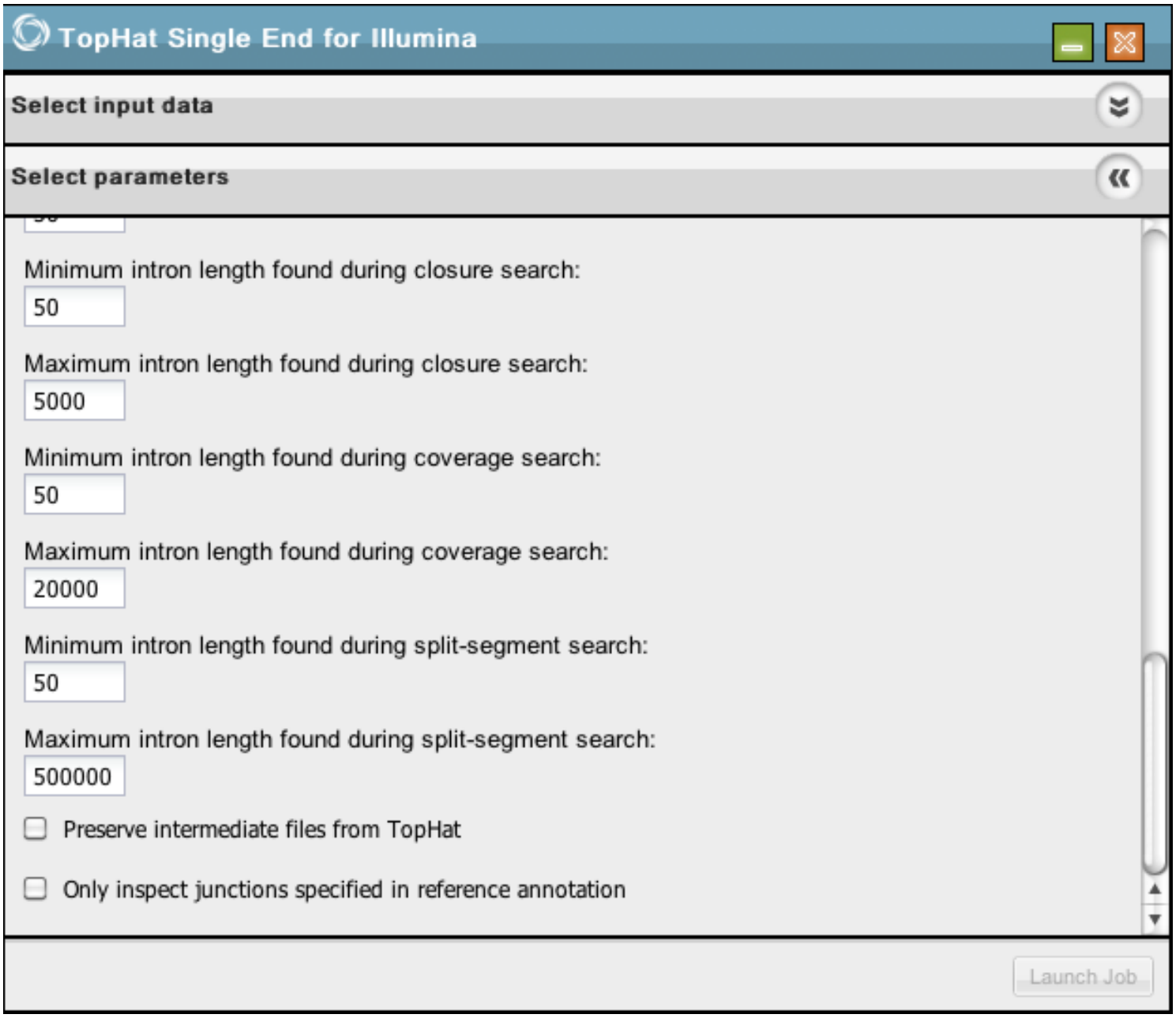


Select your desired options (continued in following images).

Select Parameters (part two)



Select Parameters (part three)


TopHat Single End for Illumina
Select input data
Select parameters
Minimum intron length found during closure search:
50
Maximum intron length found during closure search:
5000
Minimum intron length found during coverage search:
50
Maximum intron length found during coverage search:
20000
Minimum intron length found during split-segment search:
50
Maximum intron length found during split-segment search:
500000
Preserve intermediate files from TopHat
Only inspect junctions specified in reference annotation
Launch Job

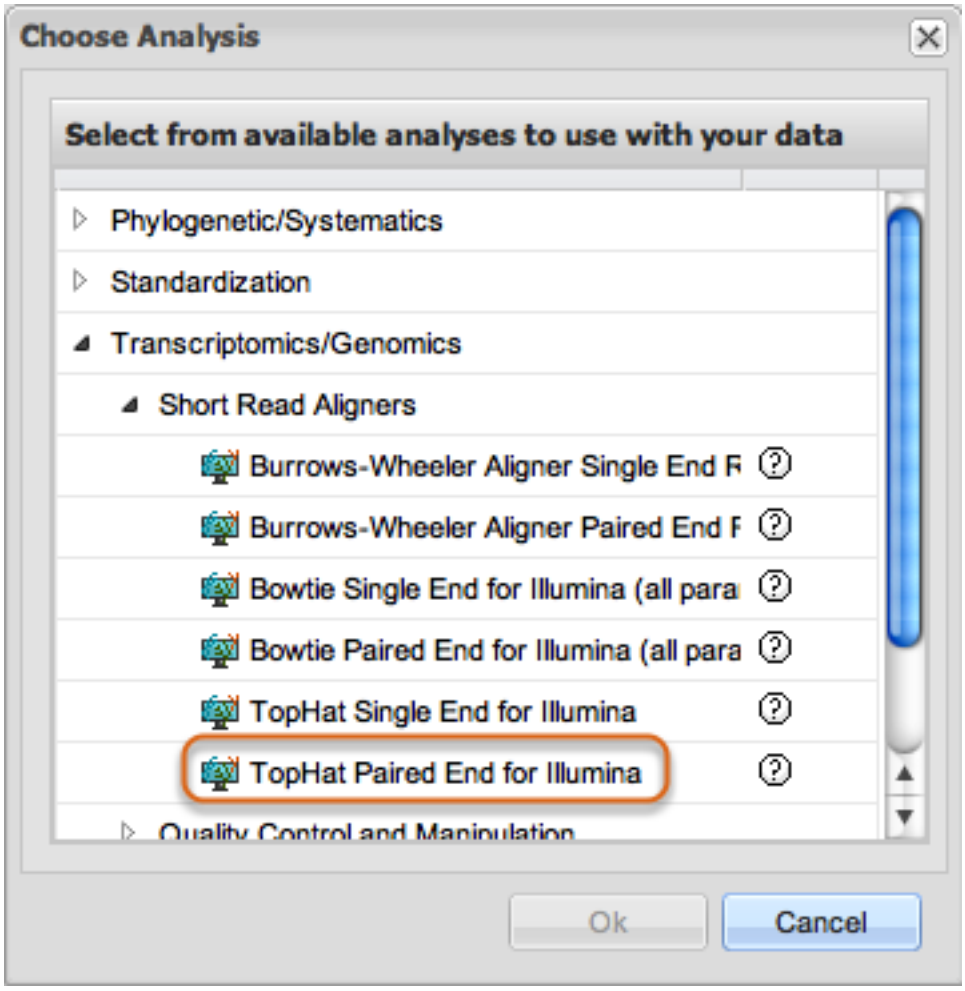
Click Launch Job.

Enter a name and description for the job and click Ok.

See Perform Analyses for information about monitoring the process and where to find your results.

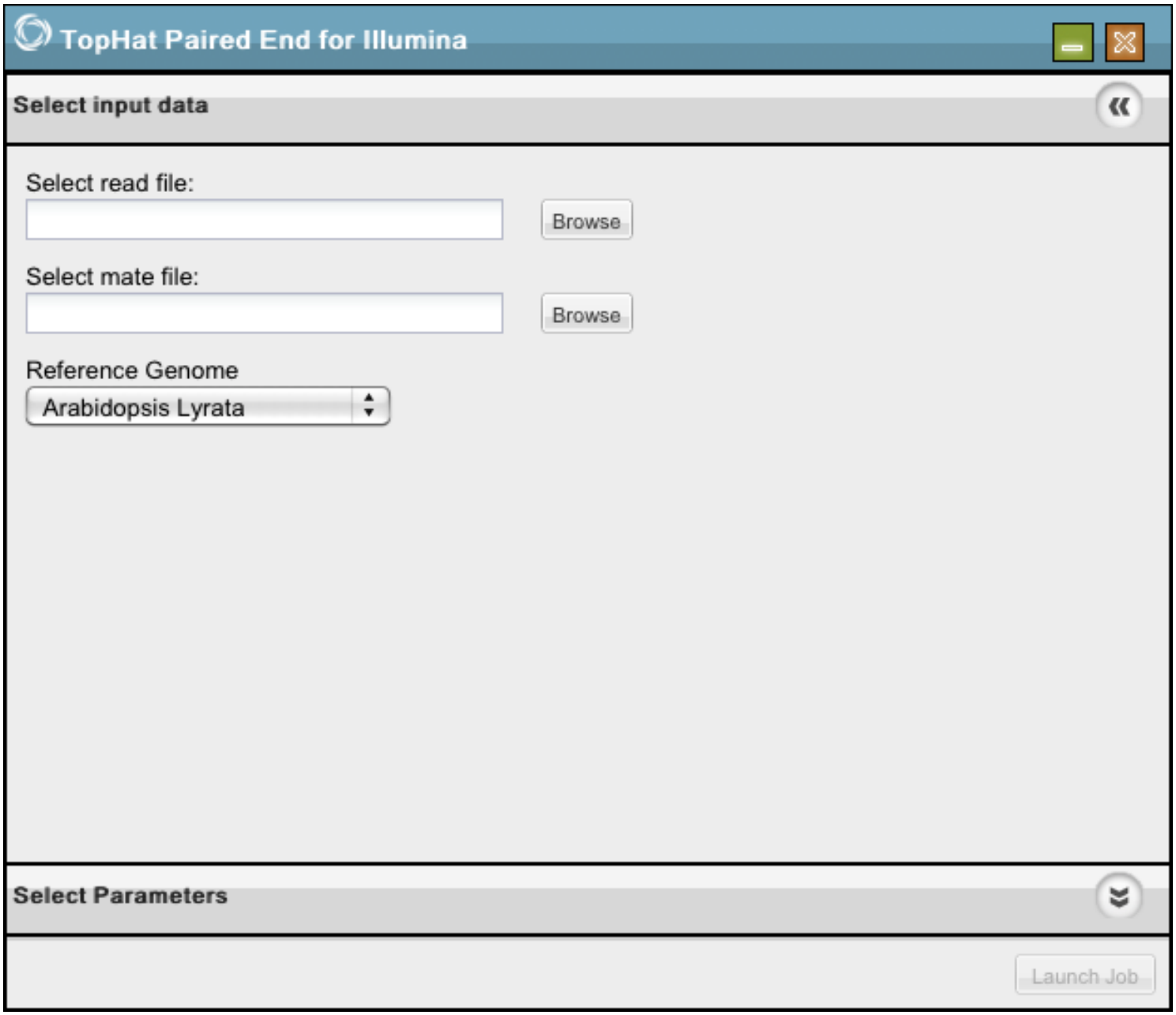
TopHat Paired End for Illumina

This analysis uses TopHat. The configuration options are set to be optimal for pair end reads derived from Illumina sequencing technology, not 454, ABI, or PacBio. A similar analysis is available for single end reads.



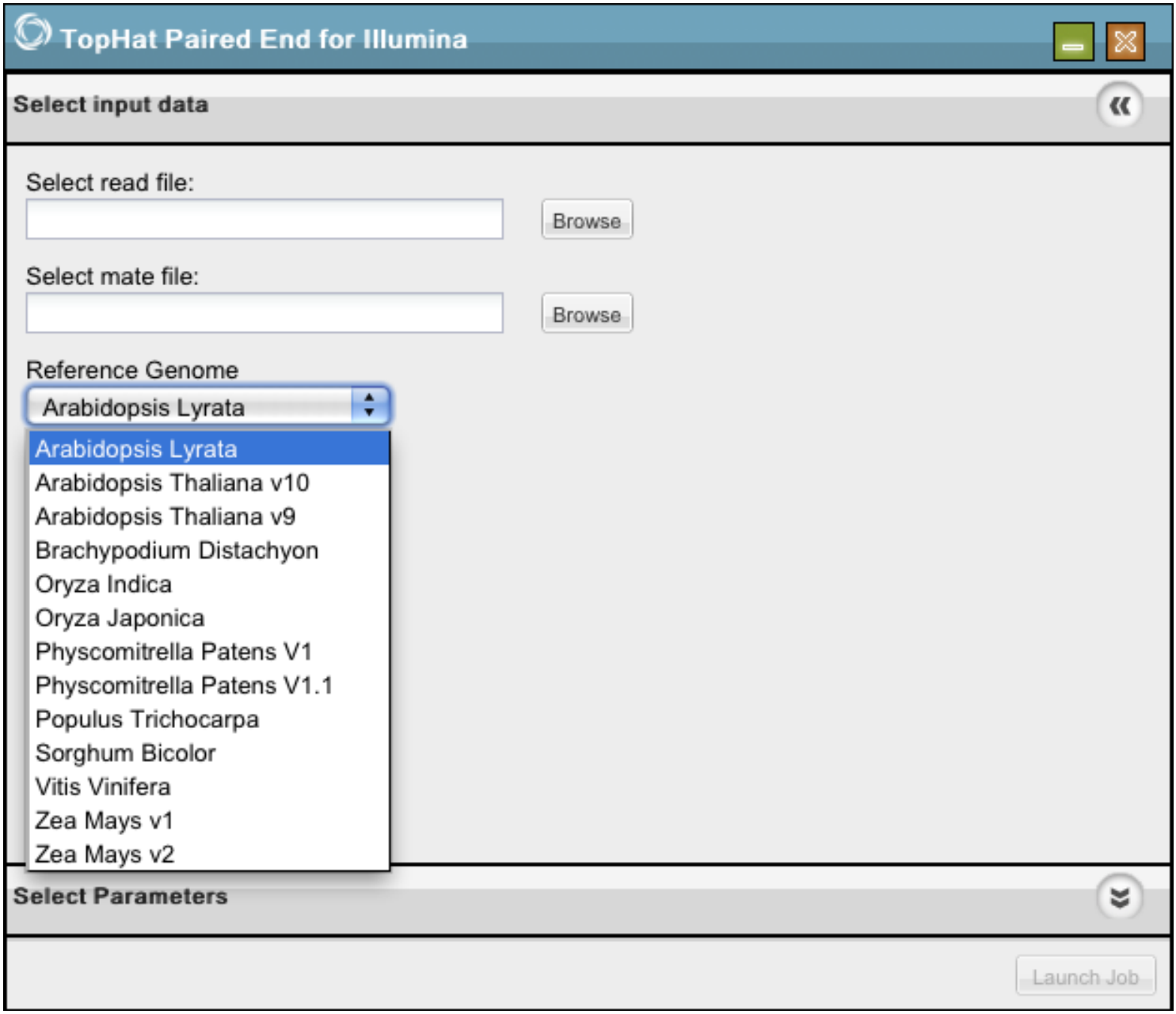
Select TopHat Paired End for Illumina from within Perform Analyses as described in that section. Click Ok.

Select input data



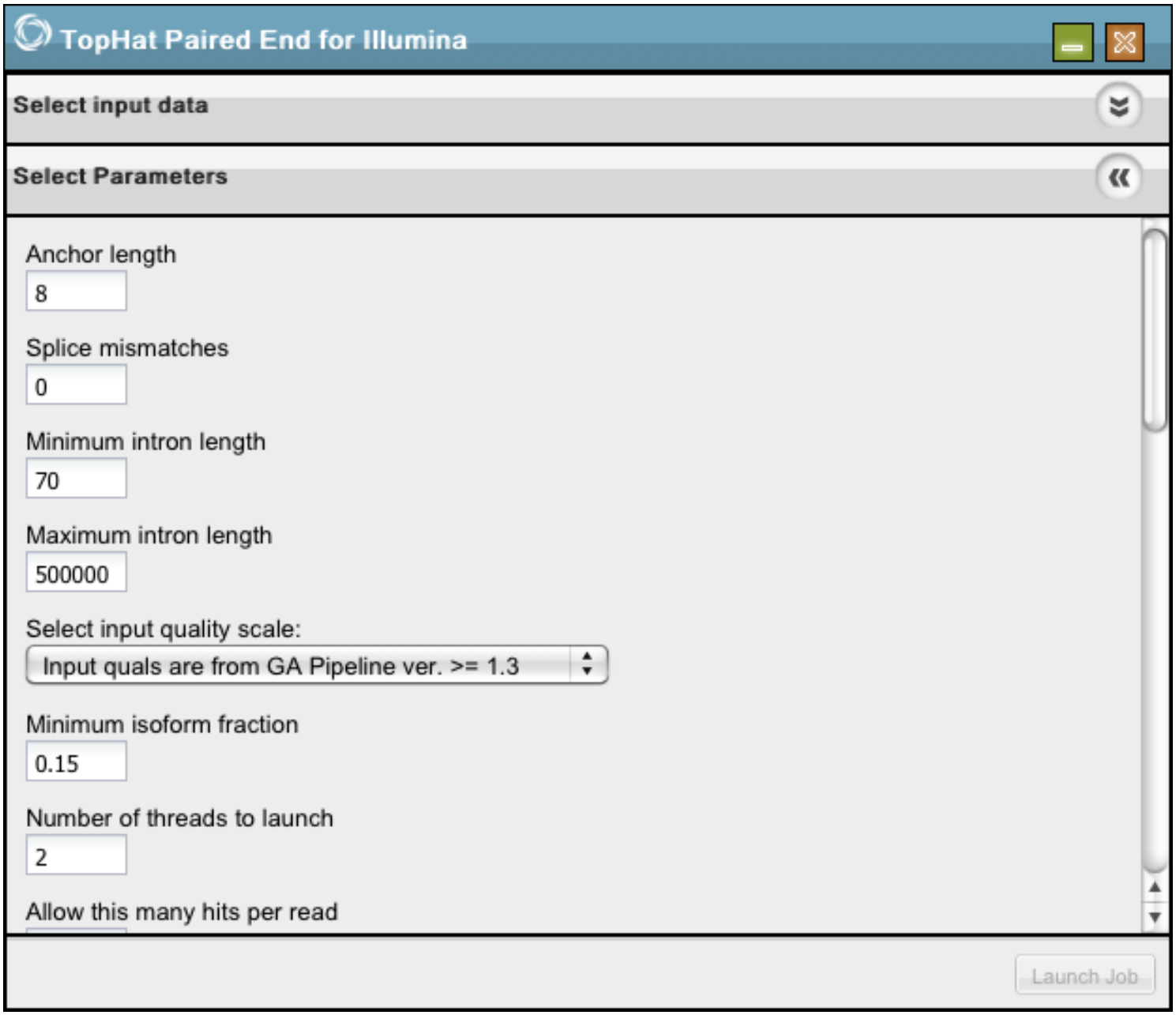
Click Add to choose the previously uploaded read and mate files you wish to align to a reference genome.

Select Reference Genome



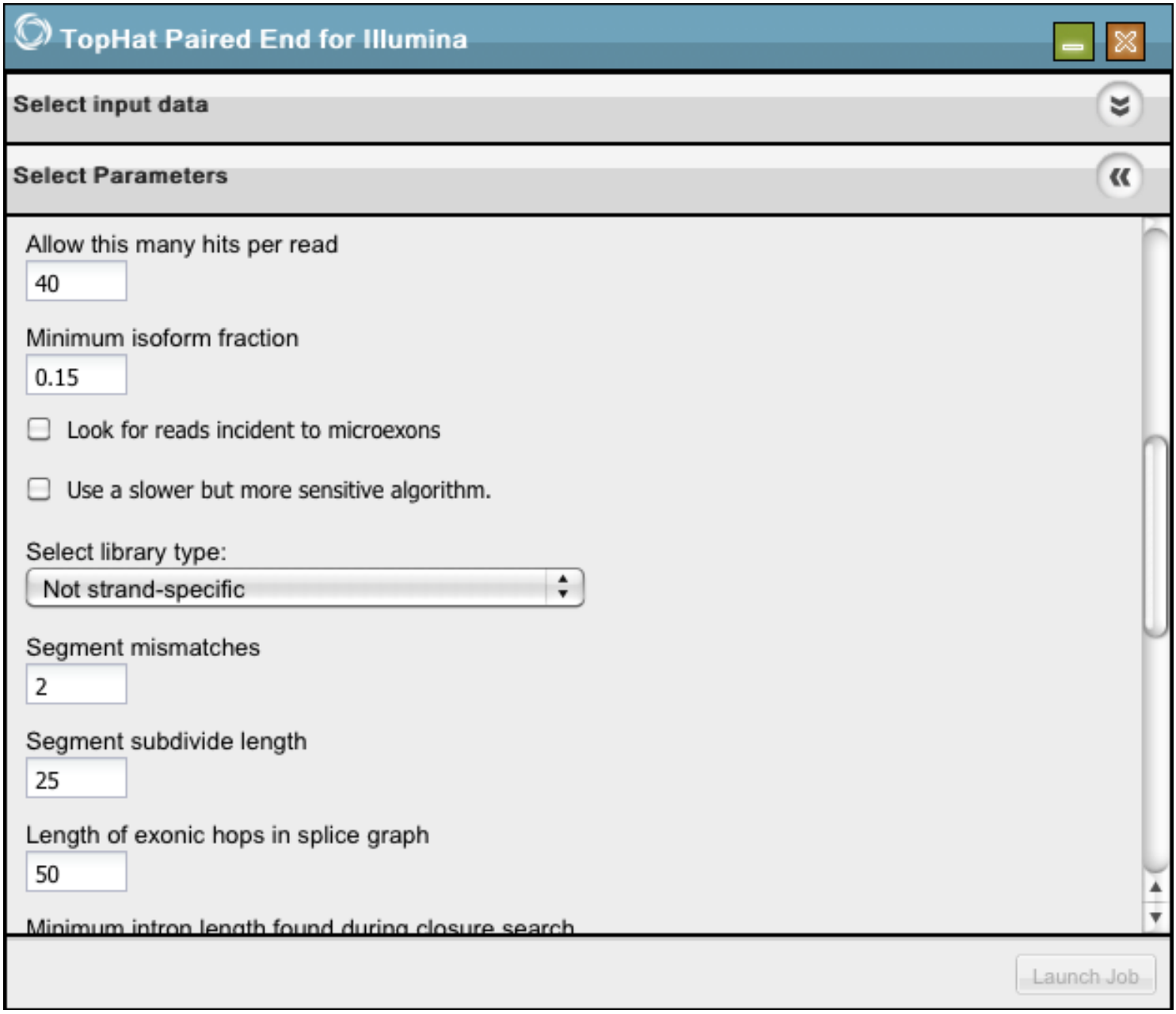
Select the reference genome.

Select Parameters (part one)

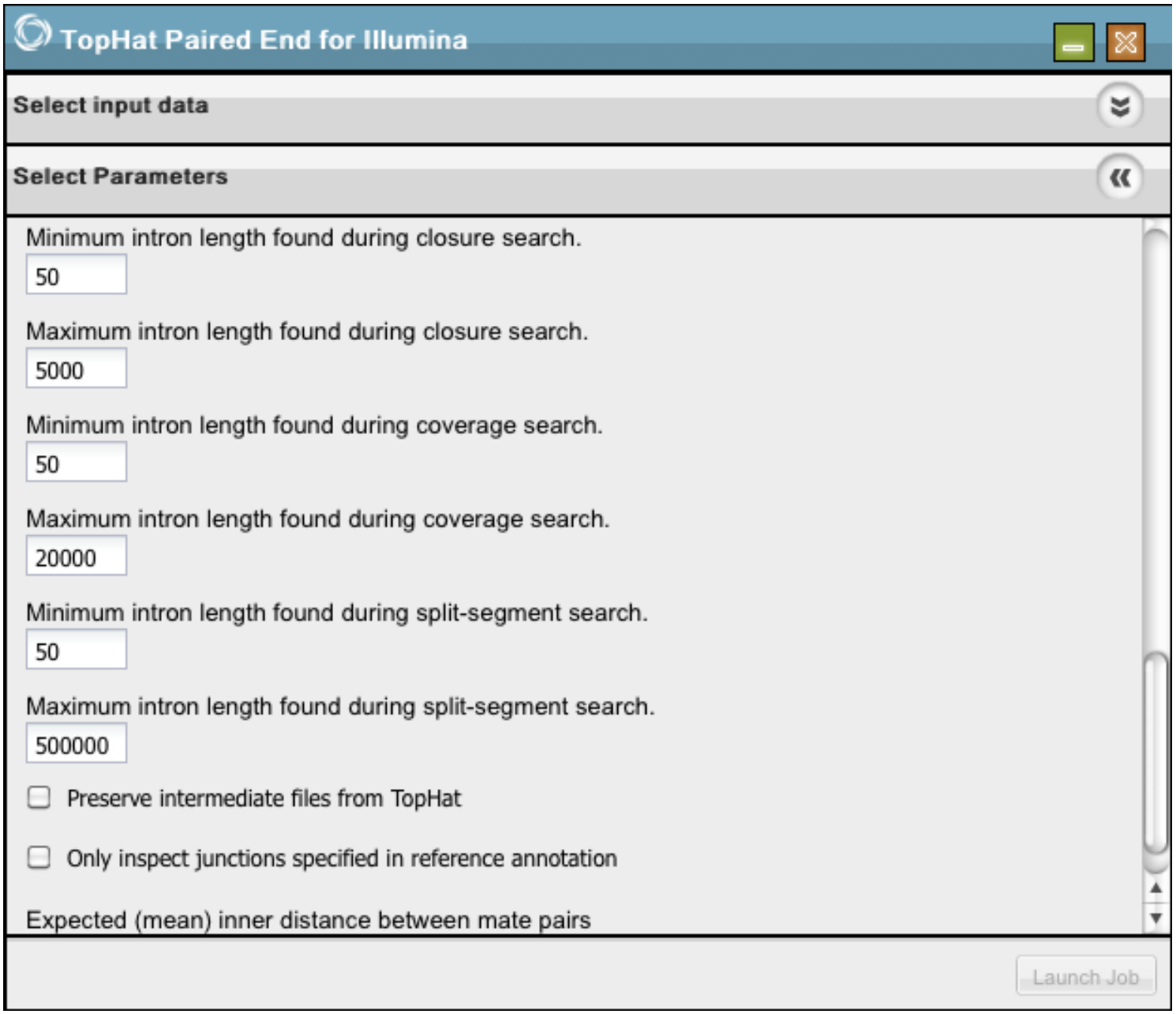


Select your desired options (continued in following images).

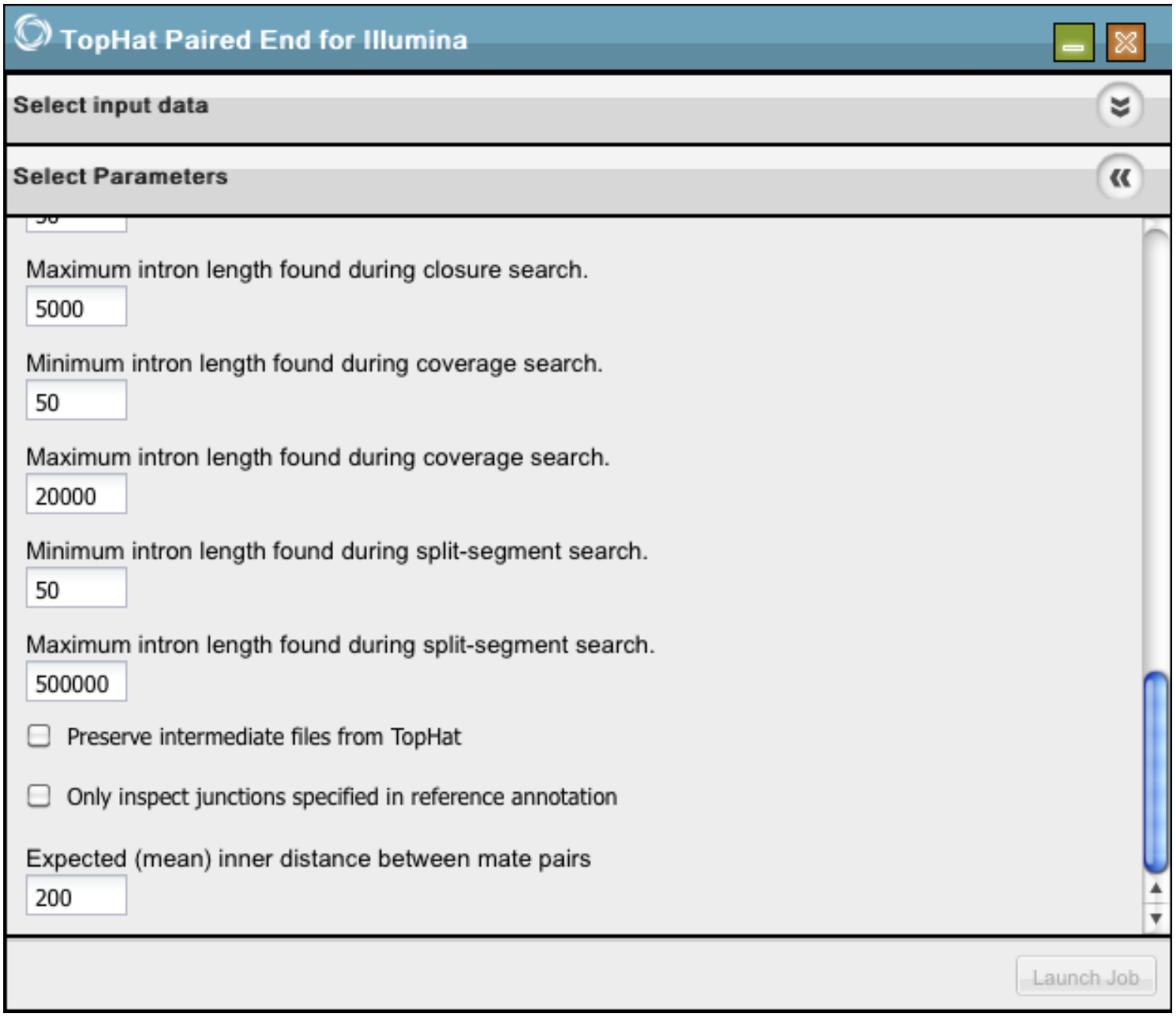
Select Parameters (part two)



Select Parameters (part three)


TopHat Paired End for Illumina
Select input data
Select Parameters
Minimum intron length found during closure search.
50
Maximum intron length found during closure search.
5000
Minimum intron length found during coverage search.
50
Maximum intron length found during coverage search.
20000
Minimum intron length found during split-segment search.
50
Maximum intron length found during split-segment search.
500000
Preserve intermediate files from TopHat
Only inspect junctions specified in reference annotation
Expected (mean) inner distance between mate pairs
Launch Job

Select Parameters (part four)


TopHat Paired End for Illumina
Select input data
Select Parameters
Maximum intron length found during closure search.
5000
Minimum intron length found during coverage search.
50
Maximum intron length found during coverage search.
20000
Minimum intron length found during split-segment search.
50
Maximum intron length found during split-segment search.
500000
Preserve intermediate files from TopHat
Only inspect junctions specified in reference annotation
Expected (mean) inner distance between mate pairs
200
Launch Job

Click Launch Job.

Enter a name and description for the job and click Ok.

See Perform Analyses for information about monitoring the process and where to find your results.

Tools

Tools Overview

Tools are software packages that perform specific tasks. We do not run tools directly in the DE; instead, we create analyses for specific uses of installed tools. An analysis may be created to use only one tool or many tools using outputs from one as inputs to another. See Tool Integration and Creating a new Analysis in the Discovery Environment for more information.

Analysis of Phylogenetics and Evolution (ape)

Analysis of Phylogenetics and Evolution (ape) provides functions for reading, writing, plotting, and manipulating phylogenetic trees, analyses of comparative data in a phylogenetic framework, analyses of diversification and macroevolution, computing distances from allelic and nucleotide data, reading nucleotide sequences, and several tools such as Mantel's test, computation of minimum spanning tree, generalized skyline plots, estimation of absolute evolutionary rates and clock-like trees using mean path lengths, non-parametric rate smoothing and penalized likelihood. Phylogeny estimation can be done with the NJ, BIONJ, and ME methods.

The above description is from http://cran.r-project.org/web/packages/ape/index.html

More information about ape is available from: http://ape.mpl.ird.fr/

ape uses the R environment.

The tool was identified for inclusion by the iPlant Tree of Life working group. The 0.3.x release of the Discovery Environment uses ape version 2.6-2.

Burrows-Wheeler Aligner (BWA)

Burrows-Wheeler Aligner (BWA) is an efficient program that aligns relatively short nucleotide sequences against a long reference sequence such as the human genome. It implements two algorithms, bwa-short and BWA-SW. The former works for query sequences shorter than 200bp and the latter for longer sequences up to around 100kbp. Both algorithms do gapped alignment. They are usually more accurate and faster on queries with low error rates.

Above description from the BWA website: http://bio-bwa.sourceforge.net/

Authors:
H. Li
R. Durban

The tool was identified for inclusion by the iPlant Genotype to Phenotype working group. The 0.3.x release of the Discovery Environment uses BWA version 0.5.9.

Contrast

Contrast compares information on the evolutionary relationships of organisms (phylogenetic trees) to test for correlated evolutionary changes in two or more traits uploaded in Newick format. Contrast reads a tree from a tree file, and a data set with continuous characters data, and produces the independent contrasts for those characters, for use in any multivariate statistics package. Contrast will also produce covariances, regressions and correlations between characters for those contrasts and can also correct for within-species sampling variation when individual phenotypes are available within a population. Contrast is a part of PHYLIP.

Above description partially from:
http://evolution.genetics.washington.edu/phylip/progs.data.cont.html

More information is available at:
http://evolution.genetics.washington.edu/phylip/
http://evolution.genetics.washington.edu/phylip/doc/contrast.html

Author:
J. Felsenstein

The tool was identified for inclusion by the iPlant Tree of Life working group. The 0.3.x release of the Discovery Environment uses PHYLIP version 3.69.

Cufflinks

Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one.

Above description from the Cufflinks website: http://cufflinks.cbcb.umd.edu/

Authors:
Cufflinks is a collaborative effort between the Laboratory for Mathematical and Computational Biology, led by Lior Pachter at UC Berkeley, Steven Salzberg's group at the University of Maryland Center for Bioinformatics and Computational Biology, and Barbara Wold's lab at Caltech.

The tool was identified for inclusion by the iPlant Genotype to Phenotype working group. The 0.3.x release of the Discovery Environment uses Cufflinks version 0.9.3.

FASTX-Toolkit

The FASTX-Toolkit is a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing. Next-Generation sequencing machines usually produce FASTA or FASTQ files, containing multiple short-reads sequences (possibly with quality information).

The main processing of such FASTA/FASTQ files is mapping (aka aligning) the sequences to reference genomes or other databases using specialized programs. Example of such mapping programs are: Blat, SHRiMP, LASTZ, MAQ and many many others.

However, it is sometimes more productive to preprocess the FASTA/FASTQ files before mapping the sequences to the genome - manipulating the sequences to produce better mapping results. The FASTX-Toolkit tools perform some of these preprocessing tasks.

Above description from the FASTX-Toolkit website: http://hannonlab.cshl.edu/fastx_toolkit/

The following are currently enabled:
• Barcode Splitter - Splits a FASTQ file containing multiple samples.
• Clipper - Removes sequencing adapters / linkers from FASTQ files.
• Groomer (Quality Rescaler) - Converts FASTQ files from Illumina 1.3+ and Solexa formats to Sanger PHRED format. This is not listed on the main FASTX-Toolkit page, but is a part of the suite. See: http://main.g2.bx.psu.edu/root?tool_id=fastq_groomer
• Quality Filter - Filters FASTQ formatted sequences based on quality.
• Trimmer - Trims (cuts) barcodes or noise from FASTQ sequences.

Author:
Hannon Lab at Cold Spring Harbor Laboratory: http://hannonlab.cshl.edu/

This tool was identified for inclusion by the iPlant Genotype to Phenotype working group. The 0.3.x release of the Discovery Environment uses FASTX-Toolkit version 0.0.13.

R Language and Environment

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment. There are some important differences, but much code written for S runs unaltered under R. R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time series analysis, classification, clustering, ...) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity. One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed.

The above description is from http://www.r-project.org/

More information about R is available from: http://www.r-project.org/

The tool was identified for inclusion by the iPlant Tree of Life working group for use with ape. The 0.3.x release of the Discovery Environment uses R version 2.12.0.

SAMtools

SAMtools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.

Above description from the SAMtools website: http://samtools.sourceforge.net/

Authors:
H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and 1000 Genome Project Data Processing Subgroup

The tool was identified for inclusion by the iPlant Genotype to Phenotype working group. The 0.3.x release of the Discovery Environment uses SAMtools version 0.1.12a.

TopHat

TopHat is a fast splice junction mapper for RNA-Seq reads using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

Above description from the TopHat website: http://tophat.cbcb.umd.edu/

Authors:
TopHat is a collaborative effort between the University of Maryland Center for Bioinformatics and Computational Biology and the University of California, Berkeley, Departments of Mathematics and Molecular and Cell Biology. It incorporates work from Cole Trapnell, Daehwan Kim, Geo Pertea, Lior Pachter, and Steven Salzberg.

The tool was identified for inclusion by the iPlant Genotype to Phenotype working group. The 0.3.x release of the Discovery Environment uses TopHat version 1.2.0.

Tree Reconciliation Demo

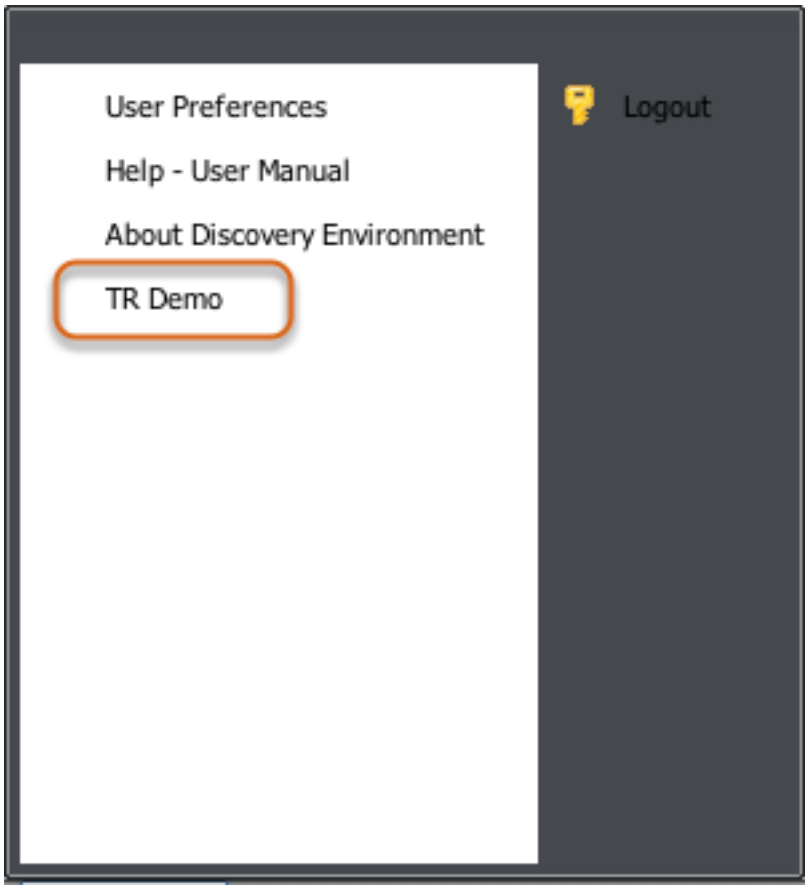
Tree Reconciliation uses an estimate of the species tree to infer the history of gene duplication and loss, lineage sorting, lateral transfer, and other events in a gene family's history.

The tool uses Muscle to align sequences, TreeBeST to build a tree, and PriMETV to display it.

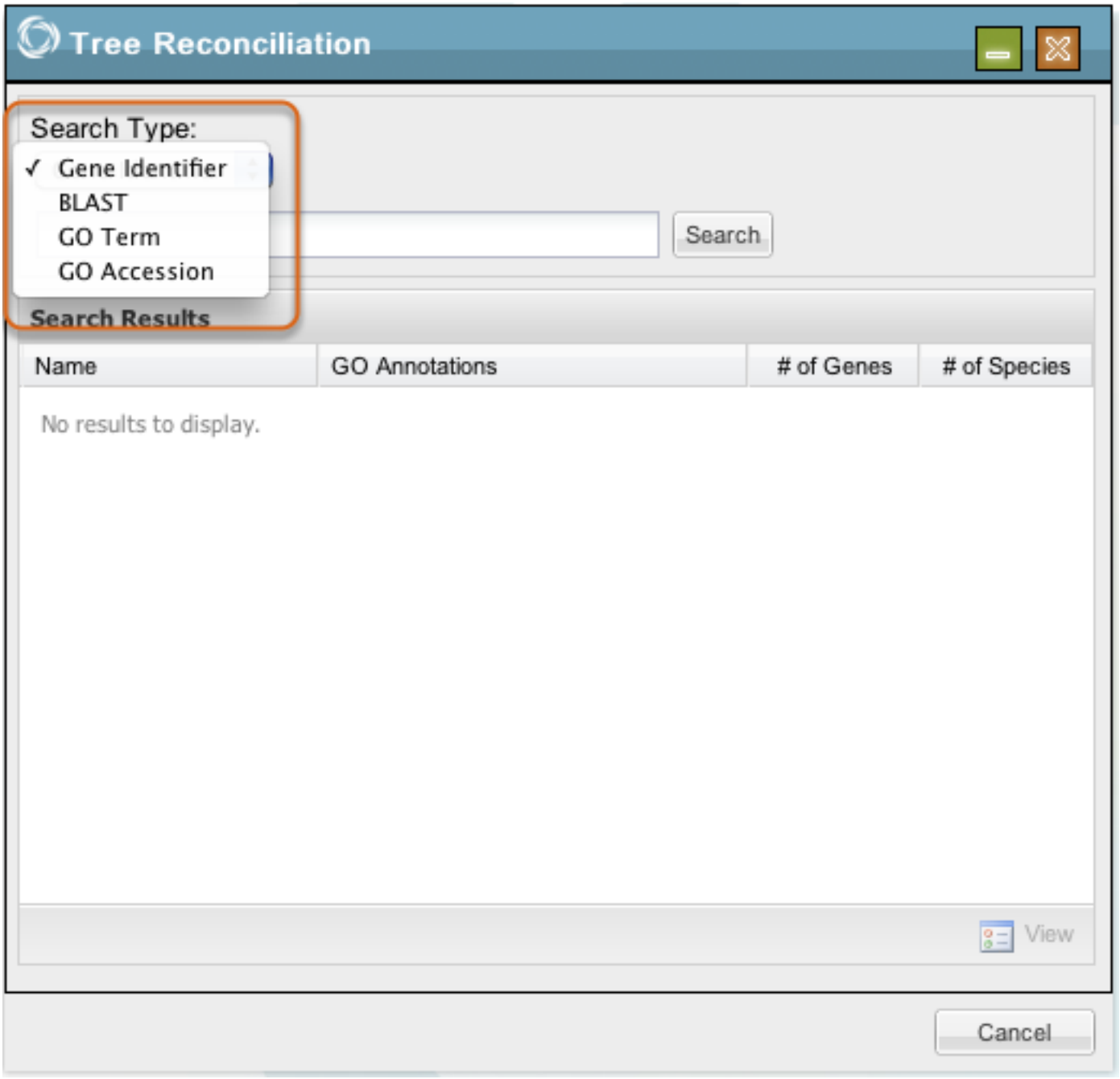
Author information for the component tools is available at each component's website, listed above.

The tool was identified for inclusion by the iPlant Tree of Life working group.

Select TR Demo from the menu



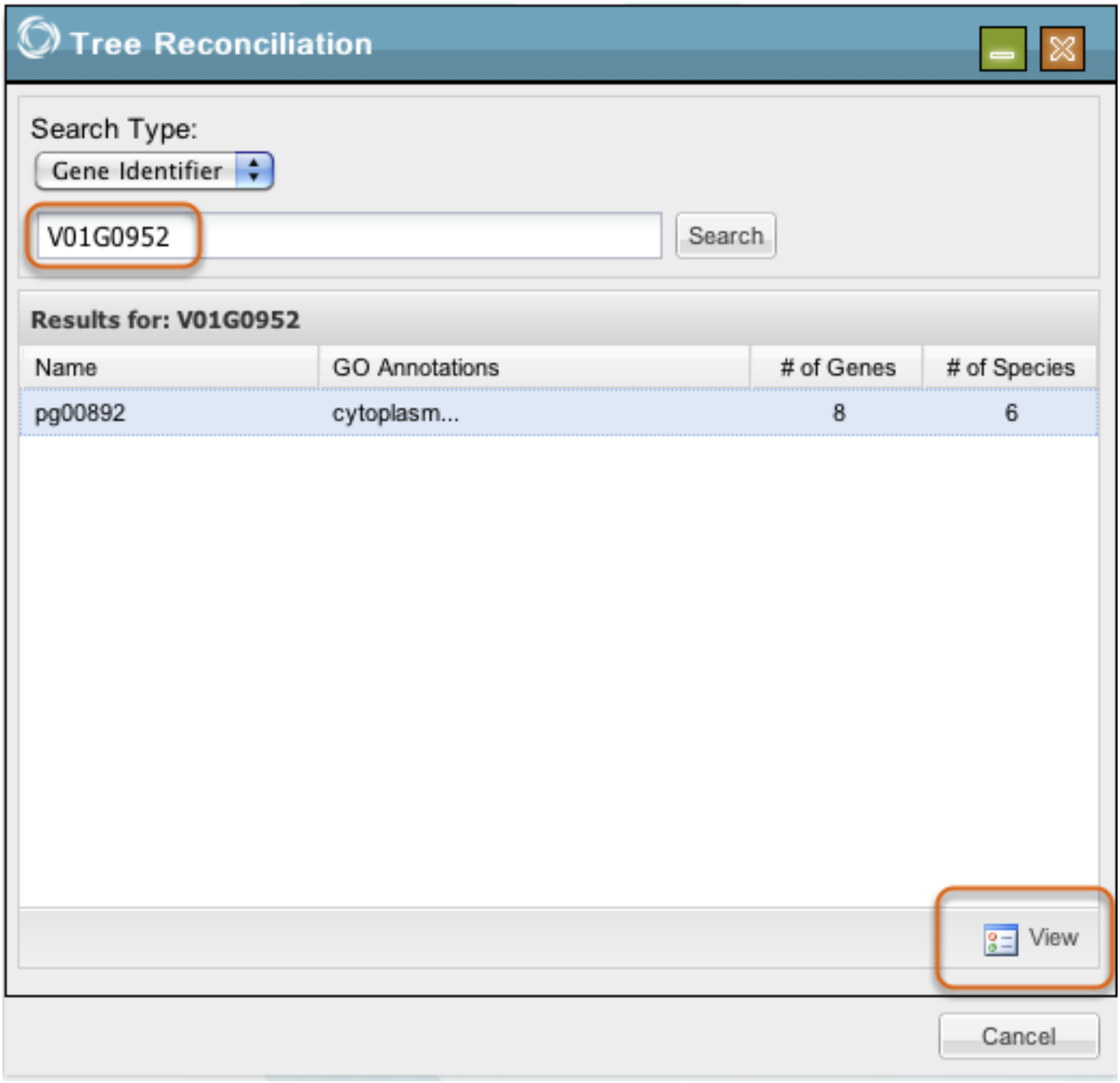
Select search type



Choose a Search Type from the drop-down box. You may search by Gene Identifier, BLAST, GO Term, or GO Accession. The genes that are currently available are Arabidopsis, Cucumber, Grape, Papaya, Poplar, and Soybean.

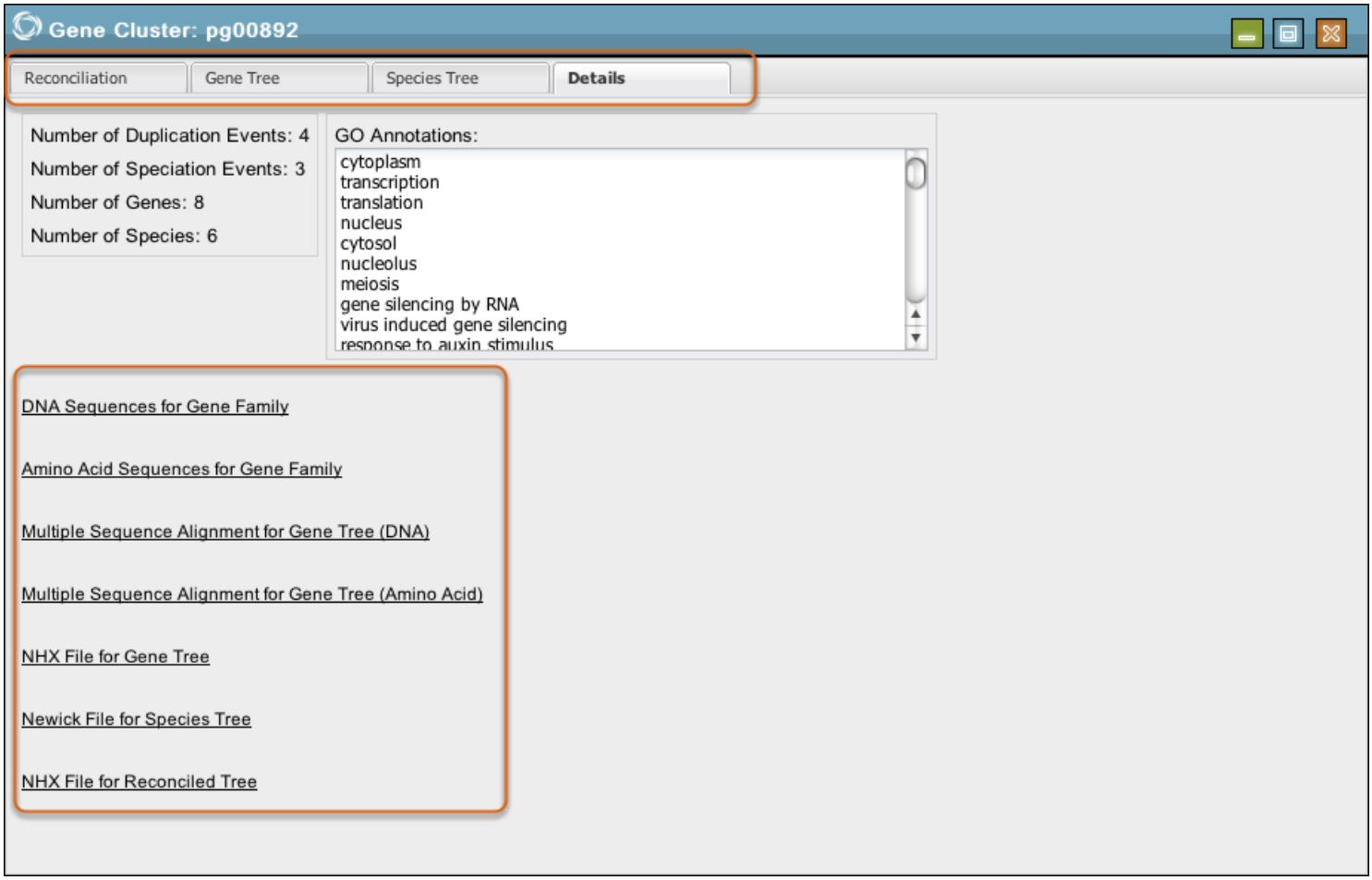
The 0.3.x release of the Discovery Environment uses BLAST version 2.2.24.

View search results



Enter your search term in the box and click Search. Highlight returned search results and click View.

View results



View and download a fat tree representation under the Reconciliation tab, a gene tree representation under the Gene Tree tab, a species tree representation under the Species Tree tab, and more details under the Details tab.

Click underlined listed items in Details to see and download the data.

Reference

Discovery Environment 0.3.0 Release Notes

This document summarizes known issues in the Discovery Environment (DE). The list is not all-inclusive, but includes the larger issues. The CORE SOFTWARE project in iPlant's JIRA has a comprehensive listing (https://pods.iplantcollaborative.org/jira/).

Each section is broken down into improvements from the 0.2.1 release to the 0.3.0 release, known issues, and future work.

This list also includes information about the Tree Reconciliation (TR) and Taxonomic Name Resolution Service (TNRS) projects. Information about Ultra High Throughput Sequencing (UHTS) and Trait Evolution (TE) are forthcoming.

Notifications

Improvements from 0.2.1

Triggered notifications

- Users can now view triggered notifications from within the View Notifications window.

Categorized notifications

- Notifications have been categorized as either "transient" or "persistent".
- Persistent notifications are related to file import or upload (success or failure) and analysis (success or failure). These appear in View Notifications and remain until a user chooses to delete them. To filter these notifications by type, users may select either the data icon, the analysis icon, or by utilize the drop down menu in View Notifications
- Transient notifications are related to file deletion, job submission, and issues where the ability to view an output (success or failure) is not available. These are presented to the user as a pop-up window in the lower right corner of the DE.

Adjustable notification display

- Notifications are displayed in descending date/time order by default, however this is adjustable. Point your cursor at the right hand of the Created Date column header will cause a down to appear. Select the arrow to choose a sorting preference from the drop down menu. Also shown is the ability to limit what columns are displayed.

Known issues

Email notifications

- There is currently an interface available to receive an email notification for long running jobs, however support services for this are not currently integrated. This issue will be addressed in a future release.

Notification persistence

- Currently, refreshing the browser will eliminate transient notifications from the main

window. This issue will be addressed in a future release.

Display

- Some of the text for the notification type may appear to be truncated. This can be fixed by moving the heading bar in the View Notifications window to allow for more room in the column.

Future plans

Additional notification types

- In a subsequent release, general notifications related to iPlant services and announcements will be added. Examples of such notifications include system downtime, community data availability, new tool/analysis capability, and others.

Icon highlighting for notification type

- Creating a feature that informs users of new notifications is being designed. Highlighting the appropriate icon to indicate job or data upload/import completion will do this. The current proposal is similar to notification behavior on Facebook, where the icons are enabled when a notification is available with a numeric representation of the number of notifications.

Email notifications

- The initial version will appear soon, but expanded and additional features are planned for future releases.

Collaboration notifications

- A new notification icon will be created for collaborations. The details of this notification type are still in the requirements gathering phase.

Analyses (jobs that are run in the DE)

Improvements from 0.2.1

0.3.0 release goal

- The goal of the 0.3.0 release was to enable submitting a job to a Condor cluster in a uniform manner for tools integrated into the DE. A service was created and hardcoded executables from 0.2.1 were re-written.

Creation of an OSM, Notification Agent, and JEX

- An Object State Management system (OSM), a Notification Agent and a Job Execution Framework (JEX) were created.

Metadata tool description

- The ability to describe tools with metadata (in JSON format) was implemented.

Flexible tool integration

- The Job Execution Framework (JEX) allows collaborators to integrate their own tools by describing the metadata (in JSON format) that is sent to the JEX and is stored by the Object State Management system (OSM). This change enables an easily repeatable process and a somewhat simple mechanism for users to integrate tools and customized implementations or uses of those tools (which we call analyses) into the DE. Core Software personnel are still needed to perform part of the process, but we have completed the first step toward making this easier for end users.

Known Issues

Progress monitoring

- This functionality is not currently available at a low level (e.g. Job is 50% complete). However, states like "running" or "completed" display in the View Analysis window for a submitted job. Low-level progress reporting is being discussed.

Job naming

- The name of the job given by the user is displayed in the View Analysis window, however the description applied to the job is not displayed. This issue will be addressed in a future release.

End date

- The user is currently not returned an "end date" completion time for the job executed. This issue will be addressed in afuture release.

Display of items in View Analysis window

- Due to the length of some of the items displayed in the View Analysis window, longer items may appear to be truncated. Users can adjust the width of column headings and view all details. Users can also maximize the View Analysis window to view these items in greater detail. Adjustments to this display are being discussed.

Display of outputs

- The user will be notified of a completed job in View Notifications, as well as via a "completed" status in the View Analysis window. The user can then select "view outputs" from the View Analysis window (or select the job name from the View Notifications window) and will be directed to the location of the outputs in the Manage Data window. These outputs will be located in generated folder that contains the name of the job and a key identifier. The key identifier is currently a large sequence. This will need to be modified to provide a user-friendly interface. This issue will be addressed in a future release.

Same name for output file and job

- The name of the output file should be the name of the job executed with "outputs" or "out" appended. This issue will be addressed in a future release.

Selection of folder for outputs

- The current workflow automatically generates a folder for outputs. Future implementations will allow a user to specify the location for those outputs. This issue will be addressed in a future release.

Job folders display

- The folder containing the outputs of analysis executions has a long name and contents are displayed in random order. To view job outputs, users can identify the correct folder by locating the folder with the name given to the user at runtime. This issue will be addressed in a future release.

Ability to stop a running job

- This functionality is not currently available. Users can remove the representation of the job from the View Analysis window, however this does not stop a running job. Consequently, outputs will be generated and displayed in the Manage Data window. This issue will be addressed in a future release.

Use of invalid file types for some analyses

- The tools integrated currently allow for some invalid file types to be selected as inputs. These analyses will execute and invalid or empty files will be generated as outputs. The fix for this issue involves changes to file handling as opposed to a "re-tooling" of the tools included in the DE. This issue will be addressed in a future release.

Display of description with outputs

- The viewer for the outputs contains a tab for the description given by the user at the time of execution. This issue will be addressed in a future release.

Performance

- Window loading and population of the window with information is not instantaneous. This issue will be addressed in a future release.

Inconsistency in the extension for outputs

- The file extensions applied to the job outputs is not consistent across tools (example: QC/preprocessing jobs will deliver different outputs, depending upon which tools are actually utilized by the user in the analysis pipeline). This is functionality is inherent in the tool itself. This issue will be addressed in a future release.

Perpetual running jobs

- There is a situation with the execution framework where communication with the monitor is lost. This will result in a job showing a status of "running" perpetually. These jobs will not complete. This issue is currently being handled and a resolution is being worked on.

Future endeavors

User customized workflows

- We will allow a user to create workflows based upon integrated tools. These workflows will be able to be generated, saved, modified and shared with groups for future analyses.

Provenance tracking

- Users will be provided a file that contains details of the analysis being executed. Included in that file will be a description of the parameters used in the analysis, data inputs and the date/time of the execution.

Default value configuration

- Users will be able to save a selected analysis with parameters that they expect to utilize on different datasets. These values may differ from the default values provided by the author of the original analysis. Users will be able to save their modified version with a name that differs from the original analysis name to indicate that it is unique to their work.

Session-based "Guest" account

- This will provide users a "preview" of the functionality that is available with a full account. It will have limits, such as no way to save work and return to retrieve it later.

Partial saving of parameters

- Users will be able to save partial entry of parameters to be used for an analysis and run at a later time.

Data management

Improvements from 0.2.1

Menu bar

- Data import and upload were obscured behind a "file" menu. This has been made more apparent to the user by exposing the functionality on a menu bar.

Data management window

- Categorization of actions a user may wish to perform on data files or folders has begun. This allows for appropriate services to be more efficiently tied to functionality and limiting the user actions to those that are appropriate for the hierarchy selected. The data management window is a work in progress. This issue will be more completely addressed in a future release.

Known issues

File movement between folders

- This functionality is currently not supported, but is a high priority on our roadmap. This issue will be addressed in a future release.

Expansion of all folders at once

- This functionality is not enabled with the current view of the Manage Data window. This topic is under discussion for integration.

Upload data from desktop

- Support for upload from a user's local environment is limited to select datatypes. Expansion of the datatypes supported is in the requirements phase. This issue will be addressed in a future release.

Import from URL

- Import from sites with a self-signed certificate fails. A fix for this is being evaluated. Import from the Sequence Read Archive is no longer supported due to a change in their format from fastq to sra (this issue will be addressed in a future release).

Display of file size

- This functionality has not yet been incorporated and is being evaluated.

Sort order of files/folder

- The display of files and folders in the Manage Data window is inconsistent and may change with each opening of the Manage Data window. A fix for this is being evaluated.

Description of files

- Users are provided the ability to create a description for their data at import. This functionality is expected in a future release of the DE. Auto detection of file types (which is the display in the description field currently) is inconsistent as well. A fix for this is being evaluated.

Filter/search

- The ability to filter or search for particular files is currently not available. This functionality is in the requirements phase of development.

File consolidation at upload/import

- Currently, users need to upload files one at a time. A fix for this is in the requirements phase.

Zipped file upload

- This functionality is in the requirements phase.

Large file deletion/upload

- This is suboptimal in the current version. Fixes are under evaluation.

Import from Phylota

- This functionality is suboptimal and improvements are in the planning stages. There are a

number of issues related to the way data is displayed as well as the general import functionality. The import from Phylota fails in the current version of the DE.

Files with duplicate names do not import

- If a file is imported that has the same name as an existing file, the import will fail. Ideally, we would add an extension to the new file's filename, such as "filename(2)".

User can not always tell to where a file will import

- Imported files are brought in to the folder currently selected by a user, however this is not always clear to the user. A note has been added to the help documentation.

Future plans

Data and file management

- Improvement to data and file management is slated for the next release of the DE. As more issues are discovered through testing of the 0.3.0 release, they will be added for evaluation for the Data Management project

Taxonomic Name Resolution Service (TNRS)

New functionality in 0.3.0

Desired name selection

- This application performs exact and fuzzy matching of a list of plant taxonomic names against a database provided by the Missouri Botanical Gardens and returns all names within a set variance. When more than one potential match is returned, the user is allowed to select the name that best reflects the intended entered name.

Selected name details

- The user is also provided links to the TROPICOS database (housed by the Missouri Botanical Gardens) for additional details. The current algorithmic pipeline includes use of the GNI parser (by Dmitry Mozzherin) and TaxaMatch (by Tony Rees).

Known issues

Matching limitations

- Current implementation only allows matching of genus and species. Work is underway to incorporate matching for full names (family through variety). A revised algorithm is needed. This issue will be addressed in a future release.

Resolving similar names

- The current implementation does not provide synonymous resolution of names. This issue will be addressed in a future release.

Entry of names

- Currently names that are entered directly into the application must NOT contain family

names. The application will not work until the GNI parser is able to accept family names.

Entry of invalid names

- The only indication a name has not matched is a return of all parts of the name in the "Unmatched" column of the application. A fix to identify the name as having "no match" is desired. This issue will be addressed in a future release.

Multiple same name return with same score

- The current version of TNRS is only performing a match to the name entered, not resolving synonyms that exist in the TROPICOS database. This information is available in the database, however at this time, all names that match the submitted name will be returned. TROPICOS does have a reference for which of these synonyms is the "accepted" name, and this name is the one that is selected as the "best match" for a user. Upon navigation to the TROPICOS web interface, this name is identified by an exclamation point ! . This issue will be addressed in a future release.

TNRS does not show results in View Analysis window

- The View Analysis window currently displays a representation of a chosen analysis and its execution for jobs that utilize the Job Execution Framework. TNRS is a web service call and does not use this framework to execute. Therefore, it does not appear in the View Analysis window. Results from an execution populate in the Manage Data window with a timestamp. This issue will be addressed in a future release.

TNRS does not use the Notifications framework

- This issue will be addressed in a future release.

TNRS job name

- The name entered is not displayed in the Manage Data window with the outputs. The user is able to identify the job only by a timestamp and a description of "Taxamatch Result". This issue will be addressed in a future release.

TNRS Manage Data window population

- The results for a TNRS job do not display consistently with other jobs and do not use the jobs execution framework. No folder is generated and the outputs return in the "root" folder for data. This issue will be addressed in a future release.

Other matching issues

- Information cannot be matched to names that are not a part of the botanical database. As a result, algae, fungi, mosses and other groups may not match appropriately in this application. We anticipate this will be fixed with updates to the database with a future release of the DE.

Download of match results

- Some browsers will request that users turn off pop-up blockers to allow for download of results from the window showing the matched names (selection of the download button). Selection to download from the Manage Data window does not present this problem, however the downloaded list from the Manage Data window is a .txt file, whereas the downloaded list from the window displaying the results is a .csv file. This issue will be addressed in a future release.

Future plans

Extending full names

- The algorithm will be extended to allow for matching for full names.

Similar names

- Synonymous name resolution will be integrated.

Additional sources

- Sources of data will be added to the database for resolution, and users will be able to specify which sources they would like to check their names against.

Tree Reconciliation (TR)

New functionality in 0.3.0

Gene family search

- This application is used to search for gene families of interest and view a reconciliation of that gene family tree with a species tree that contains those genes. For the first release of this application, a pipeline that includes MUSCLE, TreeBest and PriMETV was described.

Database search

- Users are able to search the database, which includes gene family clusters identified by John Bowers, by selecting a gene family identifier, GO term or accession or by performing a BLAST search for a gene of interest.

Search results and images

- A listing of gene families is returned that meets the search criteria and can be selected to view an image of the gene, species and fat tree representation of this data.

Download results

- Users can also download all files associated with that gene family and view a summary of the family details.

Known issues

Search interface

- The working group has redefined the items that should be available as a search parameter. A rework of this interface is underway to clarify the available options and allow

for direct selection of the family for display rather than selection of "view" to select a family.

Tree visualization

- Fat tree image - Some of the text in this image appears to be cut off. Users can scroll to get the complete image.
- Gene tree image - Curved lines and the bars for the speciation and duplication events are not standard and will be fixed when incorporation of a new tree visualization tool is implemented.
- Species tree image - Curved lines are not standard and will be fixed with incorporation of new tree visualization tools.
- Download of images - Images are not in the same format at download. A fix for this is in progress.

Saving of data from details tab

- Users do not get a notification that data is being saved. This issue will be addressed in a future release.

Interface for folder selection for saving of data

- This interface is inconsistent with the current look and feel for the Manage Data window. This issue will be addressed in a future release.

Saving of NHX files in Manage Data window

- NHX files are being identified as Nexus files upon saving in the DE. A fix for this is currently underway. Download of this file provides proper NHX format. Uploading this downloaded file in the DE will also cause the file to be interpreted as a Nexus file.

Tree visualization of saved NHX files

- The image displayed by the tree renderer in the DE that is associated with tree files cuts off text for the leaves. This will be fixed with the incorporation of new tree visualization tools. This issue exists for all tree files with lengthy names at the leaves.

Display of GO annotations

- The full annotation is being truncated. This issue will be addressed in a future release.

Search performance

- For searches that will provide a large listing of gene families (example: GO term of cytoplasm), performance is not optimal. A fix for this is being discussed.

Future plans

User capabilities

- The TR application is undergoing a complete rework to enable publication of the 1KP dataset, currently housed at TACC. Included in this rework is a basic/advanced search

interface, interactive tree functionality and a more generalized display of details for the user to make an informed decision regarding the gene family of interest.

TreeBest algorithm evaluation

- A review of the TreeBest algorithm is underway, to determine if this provides the best representation of the reconciliations. The database will also be populated with the data generated by the 1KP group as opposed to the limited subset of data that is currently available. The goal is to provide users with other data the ability to utilize the pipeline for generating reconciliations and loading this data into a uniform schema for visualizations.

Ultra High Throughput Sequencing (UHTS)

New functionality in 0.3.0

Converted and split analyses

- Many analyses that were hard-coded in 0.2.1 used multiple tools to perform extended and complex tasks. All UHTS tools were reformatted from hard-coded inclusion to instead use the new metadata format for tool integration. Then, analyses were rewritten using the new metadata format and split into discrete analyses, each focused on a specific task (often corresponding to a step in a previous analysis). This will allow for greater flexibility when user-defined multi-step analyses functionality is added in a future release.

Known issues

FASTX related analyses are currently available only for single end reads

- Paired end read analyses are planned.

Future plans

Additional tool integration and created analyses

- More tools will be integrated and basic analyses for each tool will be created.

Trait Evolution (TE)

New functionality in 0.3.0

Ancestral character estimation (ACE)

- This uses an R-based package called ape, which was installed as a tool using the new metadata method. Then analyses for both continuous and discrete versions of ACE were added to the DE using the new metadata format.

Phylogenetic Independent Contrasts (PIC)

- This analysis was hard-coded as a function in 0.2.1 and was rewritten for 0.3.0 using the new metadata methods for both tool integration and creation of analyses.

Known issues

File parsing

- Some file formats are not uploading correctly. This is being worked on currently and a fix is expected shortly.

Future plans

- being researched

Tool Integration

If you have a tool that you would like to have integrated into The iPlant Discovery Environment (DE), this can be done in just a few steps. Please contact us if you are interested in collaborating with us to do so.

The basic steps include:

* Deploying the software tool to our cyberinfrastructure
* Providing us with sample data for testing and a clear description of expected output
* Authoring metadata that tells our system about the tool and how it is used (we have samples and a clear tutorial)

Finally, to expose the tool for use, an analysis must be created. Please see Creating a new Analysis in the Discovery Environment for more information.

Please contact us if you would like to collaborate with us to integrate a tool and/or create an analysis. Please see the Tool Integration and Creating an Analysis in the Discovery Environment Quick Start guide to begin.

Creating a New Analysis in the Discovery Environment

Tools are software packages that perform specific tasks. Once tools have been integrated into the Discovery Environment, an analysis must be created. Analyses are the means by which tools are used in the DE. An analysis may include only one tool or several tools chained together into a workflow.

Tools are integrated into the Discovery Environment using a metadata description of the tool and a metadata description of the interface to that tool. All metadata is in JSON-format. Please see Tool Integration for more information.

An analysis takes a tool interface description and customizes the settings in it to suit a specific task. The analysis may choose to use all of the default values it inherits or it may set new default values, reduce parameters, or change validation criteria to define how the tool is to be used in the analysis. Analyses may be modeled for one or a combination of several tools.

Please contact us if you would like to collaborate with us to integrate a tool and/or author an analysis. Please see the Tool Integration and Creating a new Analysis in the Discovery Environment Quick Start guide on the iPlant wiki to begin.

TestData folder contents

A quick description of each of the sample data files provided in the iPlant Discovery Environment.

accepted_hits.sam

This is a SAM file produced from aligning s_8_sequence.clipper.sanger.txt to *Arabidopsis thaliana* (v9) reference genome and can be used to determine Cufflinks Transcript Quantification.

aq.trait.nex

This file contains the supporting continuous traits for the phylogenetic tree described in aq.tree.nex and can be used with aq.tree.nex for Independent Contrasts analysis.

aq.tree.nex

This file represents a 30 character phylogenetic tree that can be used with aq.trait.nex for Independent Contrasts analysis.

PDAP.trait.nex

This file contains supporting continuous traits for the phylogenetic tree described in PDAP.tree.nex and can be used with PDAP.tree.nex for Independent Contrasts analysis.

PDAP.tree.nex

This file contains a phylogenetic tree for 49 mammals that can be used with PDAP.trait.nex for Independent Contrasts analysis.

s_8_sequence.clipper.sanger.txt

This is a clipped, rescaled FASTQ file produced from removing the terminal 3' sequence adaptor from s_8_sequence.txt followed by conversion of the quality-score scale to Sanger PHRED 33 and is useful to learn and test our alignment mechanism.

s_8_sequence.txt

This is a dataset comprised of 6632564 100 bp reads from *Arabidopsis* that were generated using a protocol that may result in a terminal 3' sequence adapter and is useful to learn and test our QC preprocessing.

shorebirds.trait.nex

This file represents a set of continuous traits for the 70 bird species supported in the tree file shorebirds.tree.nex. This file can be used for an Independent Contrasts analysis.

shorebirds.tree.tex

This file represents a phylogenetic tree for 70 species of birds that can be used as inputs to an Independent Contrasts analysis

SRR026996.zmv2.sam

This is a SAM file produced from a BWA alignment of SRR026996.fastq (Mo17 genomic DNA from SRX010829) to the *Zea mays* (v2) genome and can be used for variant detection.